# Characterization of nucleosome occupancy in mammalian cells

# Characterization of nucleosome occupancy in mammalian cells

A dissertation presented

by

April D. Cook

to

The Division of Medical Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Genetics

Harvard University

Cambridge, Massachusetts

July 2014

Dissertation Advisor: Professor Robert Kingston Author: April D. Cook

# Characterization of nucleosome occupancy in mammalian cells

## Abstract

Chromatin is a complex of genomic DNA, RNA, and associated proteins. Many of the processes that occur on chromatin regulate the accessibility of the genetic material of a cell. The nucleosome is the basic subunit of chromatin, composed of a histone octamer wrapped with approximately 150bp of DNA. Alterations to chromatin structure, including to nucleosomes and their location, underlie global transcriptional diversity. A striking example of this is the so-called "open" chromatin state in pluripotent cells, characterized by loosely bound chromatin proteins and rapid nucleosome turnover, that allows transcriptional flexibility for subsequent differentiation. In contrast, differentiated cells contain compacted chromatin that can selectively block access to DNA and subsequent transcription. Thus, characterizing the physical state of chromatin is important to understanding its regulatory state.

Digestion of chromatin with micrococcal nuclease (MNase) and subsequent sequencing of the protected DNA fragments produces a map of nucleosome occupancy. Traditional MNase mapping experiments capture a snapshot of nucleosome occupancy, providing information about nucleosomes that are accessible at the level of digestion used. We analyzed regions of difference in nucleosome occupancy between embryonic stem cells (ESCs), induced pluripotent stem cells (iPSCs) and differentiated cell types using traditional MNase-seq and found that differences in pluripotent and differentiated cells are punctate and correlate with regulatory regions important for pluripotency and development. Further, our analysis shows ESCs and

iPSCs to be vastly more similar to each other in their chromatin structure than to the differentiated cells.

We then developed a new way of collecting and analyzing MNase-seq data that allows us to determine both nucleosome occupancy as well as the accessibility of DNA to regulatory factors. Our methodology discerns distinct physical states of chromatin and provides novel insights into the accessibility of regulatory regions. Additionally, we present a quantitative metric useful for characterizing local and global regions of the genome that should be useful in future cell type comparisons.

# Table of Contents

## List of Figures

## List of Tables

# Chapter 1 INTRODUCTION

Almost every cell in our body contains the same genetic blueprint: 3 billion base pairs of DNA, in duplicate, packaged into the nucleus with histone proteins. One cell's DNA, if stretched out outside the cell, would be two meters long. Wrapping the DNA around histones achieves a 6-fold compaction[1] and this beads-on-a-string structure is further folded to fit inside each nucleus. The use of the same DNA throughout the body to create diverse sets of proteins depending on a cell's type underlies our development and viability. There are hundreds of cell types in a human; each type is specialized for some function that must be created faithfully and when needed. How this diversity of cells and tissues arises and is maintained is a much studied-question.

## Chromatin, nucleosomes

Chromatin, the complex of nucleic acids and proteins residing in the nucleus of most cells of an organism, has been well-studied over the past decades. Before the structure and function of DNA was discovered it was understood that a nucleo-protein material, its components first termed 'nuclein' and 'protamin' by Meischer in the 1870, could be extracted from the nuclei of cells. The composition of this complex was determined through enzymatic microscopic and biochemical means. The name "chromatin" came about because of the staining properties of nuclear components (*chroma*=colored). In the late 1800s Walther Flemming, a German biologist, stained the contents of cells with analine dyes and he saw a well-stained string-like structure that he associated with what would be later called chromosomes. Later the roles of DNA and protein in chromatin were teased apart using virus

and bacterial transformation[2]; before these studies many believed histone proteins were encoding the heritability ascribed to this material as proteins were more complex than DNA and had a myriad of known properties[3]. Further, the structure of this nucleoprotein material was thought to be a simple model of proteins sitting on the DNA, perhaps protecting it. We now know the situation to be more complex.

DNA is a polymer of nucleotides that encodes genetic information. DNA is read by polymerase enzymes in a process called transcription to produce RNAs. The functions of RNAs are diverse: some act as messages that encode proteins, others act as enzymes, RNAs transport amino acids during translation, and RNAs can perform structural protein scaffolding functions. These processes are of the utmost importance to the cell, in that the regulation of which segment of DNA is transcribed determines the identity and function of a cell.

Chromatin is the assemblage of DNA, RNA, histone proteins and other cell-context-specific proteins that form a scaffold for nuclear processes. These processes range from DNA replication, recombination, and repair, to gene regulation [4,5]. Chromatin has along its length a basic subunit: the nucleosome. A single nucleosome is composed of four pairs of histone proteins (H2A, H2B, H3, and H4.[6,7]) wrapped with approximately 146 base pairs of DNA. Experiments using MNase to liberate chromatin fibers in the 1970s showed that different populations of chromatin sediment at different rates- newly synthesized chromatin sediments much more slowly than 'mature' chromatin. Radio-labeling pulse-chase experiments[8] were used to show that new histones are deposited at regions of DNA replication and to show the order of histone assembly. These experiments showed that the nucleosome is formed in parts,

first a H3-H4 tetramer is added to DNA behind the replication fork during DNA replication, then 2 dimers of H2A-H2B are added to complete the octamer[8,9].

This process of depositing nucleosomes as new DNA is made ensures that the genome is constantly contained in this organizational structure. Nucleosomes cover the majority of the genome. Past estimates of nucleosome occupancy of the genome were in the range of 50%-100%. These estimates were based on nuclease digestion (though not followed by positional analysis with sequencing) and polylysine binding[10-12]. Importantly it was appreciated that the chromatin preparation and the extent of digestion were important factors in making this estimation, with one scientist, AE Mirsky, stating the protein/DNA "linkage is loose, as shown by the fact that all the DNA is accessible to the enzyme; however, the wide variations in accessibility in a given complex show that the looseness of the DNA-protein link varies considerably".[12] Although the characterization of chromatin is much more complete today, this point made in the 1970s remains the same—the conditions under which you are handling chromatin and the enzyme tools you use are crucial to the conclusions you draw.

A recent genome-wide mapping effort using MNase-seq in human granulocytes showed over 99.5% of the mappable genome (the genome that is not composed of duplications or heterochromatic and telomeric repeats and thus is not in the genome assembly) is protected by nucleosomes. This coverage is distributed evenly with most nucleosome-free regions being equal to or shorter than a nucleosome[13]. It should be noted that this coverage changes with changing sequence depth. Moreover, the material sequenced represents a population of cells, so the coverage observed is the aggregate coverage and any regions that are uncovered in an

individual cell would be lost with this methodology. This suggests that the vast majority of the genome is in a nucleosome at least some of the time.

These facts taken together underscore a major question in gene regulation and other nuclear processes: if the vast majority of the genome is included in nucleosomes what does this mean for the use of DNA? During cell division DNA replication machinery must unwind the nucleosome template[14]. Nucleosomes also play a role in repair and recombination[15-17]. Perhaps the most well-studied of these processes is gene regulation. Gene regulation is heavily influenced by nucleosome positioning, as RNA polymerase cannot initiate transcription at a promoter that is organized into nucleosomes. Classic experiments in yeast show that in order to access the transcription start site, cooperating protein complexes must remodel nucleosomes[18,19]. These experiments have been followed by experiments that show the need for remodeling chromatin at transcription start sites in mammals [20,21]. However, while nucleosomes can block the initiation of transcription, RNA polymerase can read through nucleosomes when engaged in transcription. In fact, it has been reported that polymerase transcribes at the same rate whether in cells on a nucleosomal template or *in vitro* on naked DNA[1,22]. Clearly the interplay of nucleosomes and enzymatic machinery acting on DNA is complex, with nucleosomes presenting a barrier that must be overcome for cells to perform wide ranging and important processes in the cell.

## Why do nucleosomes go where they go?

Understanding where nucleosomes are, how they got there, and under what circumstances underlies an understanding of the regulation of many crucial processes, most notably transcription. Fundamentally, nucleosomes are not static; they can move along the

DNA. Additionally, histone octamers preferentially localize to some DNA sequences over others. Experiments using nucleases to probe chromatin show that nucleosomes favor G/C-rich sequences at the core or dyad of the nucleosome, and A/T-rich at the edge of a nucleosomal sequence[13,23], however no single-best nucleosome-length sequence is implicated from nuclease studies *in vivo*. To produce an optimal substrate for the production of nucleosomes for *in vitro* biochemical experiments and to determine favored positions of nucleosomes, experiments using an *in vitro* selection experiment called SELEX were carried out. $5x10^{12}$ synthesized DNA fragments were made, pooled, and used for reconstitution of nucleosome particles. Salt dialysis was used to collect the DNA fragments with highest affinity for the histone octamer, and the sequence of those fragments was determined. These experiments identified a sequence with optimal preference, now called the 601 sequence ('601' being its clone name in these SELEX experiments)[24]. In addition to this optimal nucleosome-forming sequence several "rules", or non-random features, of what constitutes a good nucleosome-forming sequence were reported. One of these rules is that TA repeats every 20 base pairs. Another is that repeating CTA at 10 or 20 base pair intervals occurs[24]. Analysis of genome content of eukaryotes shows that dinucleotide periodicity at approximately 10bp, in particular AA or TT, exists in the genome and this suggests that there is conservation of nucleosome-favoring sequence in the genome[25].

Underlying this preference of nucleosomes to a particular sequence is the fact that DNA has to bend severely to wrap 1.7 helical turns around the histone octamer to form a canonical nucleosome. There is an energy cost to this wrapping, which can be measured by approximating the different aspects of DNA deformation that must take place as it is wrapped about the histone octamer[26]. Because base composition impacts the flexibility of DNA, different

DNA templates will bend differently to accommodate the histone octamer. This information can be used to predict the occupancy of nucleosomes in the genome[27,28]. Predictions achieve better than random accuracy but do not perfect predict nucleosome occupancy. Thus, there also exist regulatory forces that supply the energy needed to push nucleosomes past the most favorable DNA sequences to the places in the genome they favor less.

Adding to the complexity of nucleosome packaging of the genome, each of the core histone proteins actually exists as a collection of isomers, for example H2A has seven variants, H3 has eight, while H4 has only one. If the histone octamer's only function were to package DNA one core histone of each type would perform this function sufficiently with no need for histone variants. For the cell to maintain and produce so many distinct histone isoforms there are likely functions important for each. For an extensive review of histone variants see Weber and Henikoff[29]. Variants carry out specific functions: centromere formation and chromosome segregation (CENPA[30]), protamine replacement through destabilization of the nucleosome in sperm cell production (TSH2B[31]), gene silencing (MACROH2A[32]) and splicing (H2A.BBd[33]). Much has yet to be studied about these variants. For example, MACROH2A was first shown to be enriched on the inactive X chromosome and important for repression of X-linked genes[32]. Recently it was found that MACRO2A on autosomes at developmental genes may be acting during differentiation by blocking a particular activating histone tail modification, H3K4me2[34]. This is a specific example of a histone variant facilitating or preserving cell identity.

Histone variants also seem to play roles in changing the structure of chromatin to make it more 'open' or 'closed' to processes that occur at the DNA. Variant H2A.Z is localized at what have been characterized as nucleosome-depleted regions (NDRs), although it is not necessary

for the depletion seen at these regions. H2A.Z is produced throughout the cell cycle, not just during replication, and H2A.Z enrichment at active genes at the NDR and gene bodies may be the result of cell cycle-independent addition of histones to DNA during transcription–mediated dimer loss[35]. H2A.Z enrichment could also be the result of a more targeted process, as there are histone chaperones that are specific to H2A.Z[36]. Combinations of variants may also behave differently, as it has been shown that H3.3/H2A.Z double variants mark NDRs as well, and may cause instability to facilitate access to regulatory regions[37]. A potential explanation for this is that these regions have a higher turnover rate and H2A.Z, which is less stable *in vivo*, helps maintain that state. More work needs to be done to explore this hypothesis, especially given that H2A.Z also binds repressed loci containing Polycomb (PRC2) in embryonic stem cells[29]. Another H2A variant, H2A.BBd, may link transcriptional processes and mRNA processing by having an impact on splicing[33]. It has also been shown that nucleosomes containing variant histones protect different lengths of DNA[33,38].

Diseases that cause alterations to histone variants *in vivo* have provided a platform for learning about how these impact cell fate[39]. For example H3.3, typically associated with actively transcribed genes, is overexpressed in esophagus cancer[40]. Other cancers that are the result of variant mutations or expression changes include melanomas chondroblastomas and gliomas, where the altered histones are variants of H2A and H3[41]. A myriad of mouse experiments and investigations into human disease further illustrate the effects of altering histone variants on development.[41]

Another histone protein, H1, is associated with the nucleosome core in a subset of genomic locations. This histone protein is sometimes referred to as the linker histone. Less is

known about the H1 histone family in mammals than the other core histone proteins. Part of the reason for this is that there are at least 10 H1 family members in humans, and knockouts of subtypes individually in mouse show limited phenotypic change[42]. H1 histone proteins have traditionally been thought to be involved in chromatin condensation and repression, though more recent studies also implicate it in 3D architecture and organization of chromatin throughout the nucleus[41,43].

The above examples shed light on the importance the make-up of the histone octamers and linker histones in chromatin and their regulation of diverse cellular processes. A nucleosome provides an important substrate for differential protein complex and chaperone binding. Nucleosomes have a fundamental impact on chromatin dynamics, and, ultimately, nucleosomes have a broad-ranging impact on chromatin integrity and gene expression.

## Layers of regulation

Individual nucleosomes can differ from the canonical nucleosome to effect regulatory change in a way that is process or context dependent. For example, core histone proteins can be exchanged for variants as described above. Additionally, a vast array of many of the histone proteins' amino acids can be chemically modified. These modifications can have an impact on nucleosome stability and nucleosome interactions with other proteins and even other nucleosomes. In particular, much emphasis has been placed on the role of histone tail modifications. Histone tails, unstructured oligopeptides at the N-termini of each of the core histones, can be chemically modified at many residues along their length. These chemical modifications include methylation, acetylation, citrulination, ubiquitination, and

phosphorylation. Reviews of what is known about histone tail marks and their functions are plentiful, see references[41,44][45-48].

Histone tail covalent modifications are commonly used to annotate the state of chromatin—they are said to mark whether a region is 'open or closed'. Open refers to chromatin that is accessible to regulatory factors, may have loosely bound proteins and may have distinct biochemical properties as well as associate with active transcription. Closed chromatin is the opposite: refractory to regulatory protein binding, and transcriptionally silent. Often these histone tail modifications correlate with cellular processes but aren't necessarily causative of those processes. H3K4me3 and K3K9ac are closely associated with active genes, H3K36me3 is indicative of transcription in gene bodies. However, in yeast, gene activation can exist in heterochromatin in the absence of marks and histone variants typically associated with active transcription[49]. In mammals H3K27me3 is deposited by the Polycomb Repressive Complex 2 (PRC2) and is associated with repressed genes in specialized cell types. Long considered a clear mark of repression, scientists were surprised to find H3K27me3 co-localized with active mark H3K4me3 at poised genes in pluripotent cells[50]. As we learn about these marks it becomes clearer that they are a useful proxy for understanding the physical state of chromatin, but are not the entity causing that state, per se.

One characteristic of several covalent histone tail modifications that is understood well is that specific domains on remodeling proteins bind these modifications. For example, some proteins harbor a bromodomain which recognizes and bind acetylated lysines[51]. An example of a protein that targets acetylated lysines on histone tails and causes a specific change in the composition of the chromatin at those locations is BRG1, a component of the SWI/SNF

remodeling complex [52]. SWI/SNF is one of at least four families of remodeling complexes, each with different potential functions in the reorganization of chromatin[53]. Relatively little is known about the vast impact the numerous chromatin remodeling complexes have in concert and what global impact they have on the architecture of chromatin. Covalent histone tail marks, while useful in combination with other data and suggestive of chromatin state, are not yet fully understood and leave the question of how chromatin state fully impacts transcription and thus cell state. As with histone variants, alterations to the amino acids in the histone tail that are substrates for modification contributes to disease, notably in the case of H3K27 methylation and cancer[54-60]. A better understanding of the mechanism of action of these marks, along with an assessment of the physical state of chromatin where they occur, will help us tease apart their causative regulatory roles.

Just as nucleosomes with modified histone tails are thought to alter binding and chromatin structure, histone core modifications impact how the histone octamer interfaces with DNA[61]. Core modifications can be made in many places, including the accessible face of the histone, the lateral surface of the histone protein, and the interface between two histone proteins. These modifications have the ability to alter a variety of characteristics of chromatin including histone-DNA contacts and nucleosome stability[62-64].

## How packaging affects transcription and cell fate

When DNA wraps around the histone octamer the ability of DNA-binding proteins to bind fundamentally changes, changing the transcription potential of the locus. Either the DNA is contacting the histone octamer, occluding protein binding, the DNA is bent severely enough that the recognition site for the protein is not recognizable, or the DNA is bent in a way (in

combination with its placement on the histones) that creates or retains a recognizable binding site. Whichever the case, a nucleosome presents a regulatory landscape much more complicated than DNA and its binding proteins alone.

Transcription occurs on a nucleosomal template. RNA polymerase II (RNAPII) must be able to access DNA. It has been shown *in vitro* that nucleosomes occlude polymerase binding, but it is also the case that transcription can process through a nucleosomal template *in vivo*[1]. In fact, nucleosomes in different positions with respect to the transcription start site have differing abilities to impact the rate of transcription. The +1 nucleosome (just downstream of the TSS) seems to be a barrier for RNAPII, with the extent of occupancy correlating with how severe a RNAPII stalling event is. In contrast, the nucleosomes in the body of a gene being transcribed are not thought to have as strong an effect on RNAPII passing[65]. Further, it appears that the composition of the nucleosome histone octamer has an impact on how easily the RNAPII can pass with H2AZ-containing nucleosomes presenting less of a barrier to RNAPII than non-H2AZ-containing ones[65].

Despite a widely held belief that nucleosomes are generally an impediment to protein binding and RNAPII passing and are thus a 'block' to changes in transcriptional activity[18,20], some reports suggest that nucleosomes themselves can also be a substrate for transcription factor binding[66-70]. For example, both p53 and progesterone receptor have well-defined DNA binding sites in the mammalian genome. Recent studies showed that nucleosome occupancy maps include nucleosomes at the binding sites of these factors before stimulation and binding of the factors. The nucleosome occupancy then diminishes at these binding site following a system-specific stimulation. The pioneer transcription factor FoxA has been extensively studied

and has been shown to co-occupy DNA with a core histone protein[70,71]. It is unclear exactly what the role of FoxA1 and FoxA2 are on nucleosome occupancy globally, as it has been shown both that 1) these proteins are not responsible for nucleosome occupancy through knock out and MNase-seq[72], and 2) that these proteins exert a direct effect on nucleosome occupancy[73]. It is likely that experimental methodology plays an important role in this discrepancy. These results taken together suggest that nucleosomes play a role in the binding and activity of transcriptional activators, and show that much is left to be understood.

## Chromatin is dynamic

As mentioned above, the histone octamer is not fixed in position along the DNA, it can move along the DNA and come on and off the DNA partially or fully. Additionally the DNA wrapped around the histone octamer is thought to be dynamic, as well, unwrapping and rewrapping, or breathing, allowing access to nucleic acid and protein as it does so. This unwrapping and rewrapping allows transient access to DNA binding proteins, including transcription factors. What fraction of factors binds to DNA during these unwrapping events and which factors actively bind and/or displace nucleosomes remains an open question. The foundation of answering these questions lies in characterizing this breathing. Both equilibrium constants and rate of breathing have been experimentally derived. Regions of DNA at the edge of wrapped DNA tend to move further apart and stay apart longer than DNA located near the middle of the nucleosome[74]. The rate of unwrapping is about every 250ms with re-wrapping occurring on a much shorter time scale[75].

In addition to nucleosome 'breathing', nucleosome turnover, or the loss and replacement of histones on DNA, impacts DNA accessibility. Histone variants and modifications

as well as other binding proteins and DNA modification have an impact on how strongly the nucleosome is held together, and ultimately how often a whole nucleosome dissociates and histones become free of DNA. Nucleosomes can be formed in a replication-dependent and replication-independent manner. A variety of protein complexes are responsible for assembly of nucleosomes under a variety of conditions. One way of investigating turnover is with a system of tagged histone proteins. For example, to study histone variant, H3.3, which is deposited exclusively in a replication-independent manner, tagged H3.3 protein was induced and measured at subsequent time points. This study found that there are several rates at which the H3.3 protein is incorporated into chromatin that are correlated with the activity status of the region[76].

Chromatin is dynamic in itself, with constant breathing and turnover that is influenced by a myriad of histone protein variants and modifications. Moreover, forces external to chromatin play an important role in its composition. There exists an extensive array of proteins and protein complexes whose role is to remodel chromatin. There are at least 4 families of these proteins that have been evolutionarily conserved, SWI/SNF, ISWI, INO80/SWR1, and CHD, and they use energy from ATP to move, eject, order, space and assemble nucleosomes along chromatin in a sequence-independent manner[53]. These complexes have been studied extensively *in vitro* in biochemical assays. For example the ISWI protein can slide a nucleosome from the middle of a piece of DNA to the ends when acting alone, but actually moves nucleosomes towards the middle of a fragment of sequence when associated with ACF or CHRAC complexes[53]. Also, much is known about how remodelers are recruited *in vivo* during cellular processes, which is largely influenced by accessory proteins as well[77].

Chromatin remodelers and their complexes are fundamentally important in many processes, and because of this their disruption results in broad ranging diseases[78]. For example SWI/SNF is involved in the control of growth, and can cause developmental defects in mouse studies. Along these lines it is not surprising that the SMARCB1 (SNF5/INI) component of the BAF (human SWI/SNF) complex is inactivated in cancer[78]. SWI/SNF family member mutations in mouse studies usually have lethal effects, but there are examples of human developmental syndromes that are caused by mutations in remodeling proteins as well, such as CHD7 mutations causing CHARGE syndrome[79].

A good deal is understood about remodeling protein action *in vitro*, and mutations in these remodelers underlie a variety of diseases. However an assessment of their effect on the physical state of chromatin in a particular biological context *in vivo* is still largely elusive.

## Ways of looking at chromatin architecture

Currently, methods for probing how chromatin is organized in the mammalian nucleus range from microscopy (FRET, immunostaining) to biochemistry[80] with little in between. NOMe-seq uses a methyltransferase to mark all GpC nucleotides that are not already methylated or protected by a nucleosome, followed by bisulfite sequencing, providing a map of both nucleosome occupancy and methylation. This method is dependent on both methyltransferase activity and GC content in any given locus analyzed. A number of methods have been developed that harness the ability of nucleases or chemicals to cut (MNase, DNaseI), or sonication to shear DNA (FAIRE). After cleaving exposed DNA, the remaining fragments can be sequenced, providing a map of protected fragments of DNA. FAIRE, (formaldehyde assisted isolation of regulatory elements) is, as its name suggests, only relevant in a cross-linked setting,

while nucleases can be used in native or cross-linked samples. Further, FAIRE-seq methodology physically separates the DNA that is not protected in nucleosomes and sequences it, the outcome of which could be susceptible to level of sonication. FAIRE has been shown to produce maps of open chromatin similar to those produced by DNaseI[81]. MNase maps regions that are open similarly to DNaseI and FAIRE, as well as mapping the occupancy level of nucleosomes in chromatin. An interesting new technique called ATAC-seq uses transposons to directly profile accessibility in active chromatin by inserting sequencing adaptors into the genome of a cell.[82] This assay is able to map accessible and inaccessible (nucleosome) regions through bioinformatic analysis of resulting sequence fragments, however because the transposase has a preference for active regions and cannot map condensed regions, this method is limited in scope. While MNase has been reported to have sequence specificity, another nuclease, caspase-activated DNase (CAD), can be used to probe accessible nucleic acid in the cell and a comparison of the resulting maps of CAD and MNase-digested chromatin showed no major differences, suggesting MNase cutting bias does not significantly effect the resulting nucleosome occupancy map.[83]

Historically, the most common way to probe nucleosomes in chromatin is through the use of MNase. Because of its ability to digest any linker DNA across the genome with some, but not a great deal, of cutting bias, as well as the ease of carrying out an enzymatic digestion, it has remained a useful tool for years. Much has been learned about the architecture of local chromatin through the use of MNase, and with the dawn of the age of genome-wide sequencing we have revisited this powerful tool and it is the centerpiece of this report of chromatin architecture.

## MNase

Micrococcal nuclease, or MNase, has been an integral tool in probing the contents of the nucleus for decades. A endo-exonuclease originating in Staphylococcus aureus, micrococcal nuclease activity was observed in the 1950s and the enzyme was isolated and characterized in the 1960s[84]. Early use of MNase centered on determining DNA nucleic acid content. Later it was found that MNase digestion of nuclei and isolation of nucleic acid with subsequent agarose gel electrophoresis shows a characteristic ladder pattern[85], with repeating units of approximately 200bp[7]. This ladder phenomenon, along with other data, was synthesized into a basic understanding of chromatin structure[7], and this understanding has been growing for over 50 years.

The specific activities of MNase have been well studied. MNase preferentially cleaves RNA over ssDNA, and ssDNA over dsDNA. MNase surrounds the DNA to cleave between bases, much like a bolt cutter, explaining why it is less effective at cutting DNA lying atop the histone octamer in a nucleosome versus free DNA. If digested extensively with MNase, the 150 bp of DNA in a nucleosome will be further cleaved in predicable steps, then MNase will digest all DNA down to 3´ phosphomononucleotides and dinucleotides. MNase has been long known known to have sequence specificity[86,87]. It has been noted that the pattern of protection in MNase-digested naked DNA is similar to that of chromatin, lending doubt to the validity of MNase profiling of chromatin structure[88]. Further, MNase-seq data is enriched for fragments with a skewed nucleotide content, raising the question of whether MNase sequence-bias is a significant factor in MNase-seq maps[89]. A direct comparison of MNase with a nuclease with a distinct mode of cutting, caspase-activated DNase (CAD), showed no significant bias to

nucleosome mapping[83]. Additionally, a study using extensive sonication followed by H3 ChIP showed that the characteristic nucleosome free region at the transcription start site (TSS) was present in their data, despite concerns that the TSS sequence itself is more susceptible to MNase cleaving[90].

Several studies in yeast have illuminated the importance of treating MNase as an enzyme that liberates different populations of DNA fragments with changing concentration.
It was stated in Rizzo et al. that technical differences in MNase conditions cause major differences in the outcome of MNase experiments. They simulate the effect of different extents of digestion and suggest a method of matching samples that reduces variability between replicates[91]. In a model of actively growing yeast, Weiner et al. showed that there exist "sensitive nucleosomes" within "nucleosome-free regions"[92]. Said differently, previously reported nucleosome-free regions are not truly nucleosome free, but are likely associated with nucleosomes lost in conventional MNase digestions[92]. Xi et al. described the occurrence of "fragile nucleosomes". An MNase digestion releasing approximately 10% of chromatin and a complete digestion were performed and sequenced separately; this showed that a subset of regions thought to be nucleosome-free had nucleosomes that were quickly liberated and then lost in more extensively digested samples[93]. This makes sense when considered with nucleosome turnover, described above. It is also well known that active chromatin (chromatin associated with transcription) has different biochemical characteristics than closed, and this property was taken advantage of by Teves and Henikoff when they developed a method to profile active chromatin by MNase digestion following salt fractionation[94]. Given that the majority of MNase-based experiments in mammals still use one digestion condition, these

studies above suggest that the use of MNase as a tool to determine nucleosome architecture is still evolving with new insights yet to be found.

## Recent studies using MNase as a tool in mammals

Nucleosomes and where they exist in chromatin play a regulatory role in transcription and thus cell fate. Many proteins and protein complexes have been conserved over time whose role is to move nucleosomes and remodel chromatin. Nucleosomes have a tendency to move towards or form at 'favored' sequences, so any deviation away from such sequences represents the outcome of a regulatory event. Discerning where histone proteins in distinct cell types differentially protect DNA should provide a map that can be used to locate regions potentially under regulatory control. Further, harnessing the ability of MNase to liberate different populations of fragments of DNA using different levels of enzyme should allow for probing the chromatin structure in ways not done before.

### General nucleosome map features in eukaryotes

Through the use of MNase, much has been learned about yeast nucleosome occupancy, partly because it has a relatively small genome for a eukaryote making it amenable to sequencing methods. For example, when nucleosome occupancy is averaged around transcription start sites in yeast a characteristic pattern emerges: a pronounced nucleosome-free region (NFR) at the TSS and a strongly positioned nucleosome upstream (-1) and downstream (+1) of the TSS, with phased nucleosomes emanating out from the TSS[95]. Yeast nucleosome occupancy is also characterized by 'barrier', or 'statistical', positioning.[96-98] This phenomenon can be described as some barrier setting up an array of evenly spaced nucleosomes, usually a sequence-determined nucleosome free region (NFR) or sequence-

specific protein bound to DNA. It is not clear what causes this phasing. It has been postulated that only a barrier (a bound protein or specific DNA sequence) is required, and nucleosomes, which can move freely, will adopt a phased state around the barrier. Furthermore, in yeast specific nucleosome changes can be observed in situations where cells undergo perturbation, for example heat shock[99]. Nucleosomes have been seen to be evicted at promoters being activated, and gained at genes being repressed. Notably not all changes in nucleosome occupancy in yeast are associated with a transcriptional change[99]. Yeast data suggests nucleosomes are fairly well-positioned and change position in meaningful ways.

Questions remain as to what extent nucleosome occupancy changes in different cells across species and within an organism. Until recently, it was unknown what the global landscape of nucleosome occupancy is in mammalian cells as it was prohibitively expensive to perform MNase digestion followed by whole-genome profiling (whether tiling array or sequencing). With the dropping price of whole-genome sequencing came genome-wide maps of nucleosome occupancy in mammals[13,52,72,73,100-105]. Fundamental characteristics of nucleosome occupancy were examined and described in these studies. Further attempts to compare cell types or states were made and the description and outcome of those studies is summarized in Table 1.1 and in the text below.

Table 1.1
Representative Mammalian MNase-seq studies

| first author | PMID | year published | sequencing /detection method | organism | biological sample | # cells/ amount of chromatin | native or XL? | Mnase methodology | size selection? |
|---|---|---|---|---|---|---|---|---|---|
| Schones | 18329373 | 2008 | SE seq | human | resting and active CD4+ T cells | unk | native | MNase to generate approximately 80% mononucleosomes and 20% dinucleosomes | agarose gel (150bp) |
| Valouev | 21602827 | 2011 | SE seq | human | granulocytes, CD4+ T cells, CD8+ T cells | unk | native | cells snap-frozen and crushed to release chromatin, followed by micrococcal nuclease treatment. | agarose gel (120–180 bp) |
| Li | 21623366 | 2011 | PE seq | mouse | Foxa1- and Foxa2-deficient liver as well as mouse embryonic stem cells and mouse heart | unk (4 livers/sample) | native | chromatin released from nuclei in buffer containing 0.1 N CaCl2, partial digestion = MNase 15 min, and full digestion = 30 min. pooled partial and full digestion of four mouse livers; extended 36-nt reads to 120 bp according to the average size of nucleosomal DNA fragments | "Mononucleosomal DNA was collected" |
| Gaffney | 23166509 | 2012 | PE seq | human | 7 lyphoblastoid lines | unk | native | samples in ice cold buffer transitioned to 37 deg for 12 minutes | 3.3% Nusieve agarose gel (147 bp fragments); discarded fragments of extreme size (outside the central 95% range of 126–184 bp) |
| Kundaje | 22955985 | 2012 | SE seq | human | K562, GM12878 cell lines | unk | native | same as valouev | agarose gel (120–180 bp) |
| Teif | 23085715 | 2012 | PE seq (at least 50bp) | mouse | ESCs, NPCs, MEFs | unk | native | cells harvested, resuspended in low-salt buffer at 4 °C. Digested with 0.5 units Mnase/microliter 6–11 min at 37 °C. MNase digestion was stopped by putting the samples on ice and adding EDTA to a concentration of 10 mM. | 2% egel; fragments corresponding to mononucleosomes or dinucleosomes |
| Hu | 21795385 | 2011 | SE seq | human | HSCs, CD36+ cells | unk | native | same as Schones et al | agarose gel (120–180 bp) |
| Tolstorukov | 23723349 | 2013 | helicos, PE-seq | mouse | primary mouse cells with Snf5 or Brg1 deleted | MEFs from multiple embryos were pooled: WT, 9.4 × 10^7 cells were digested; Snf5, 5.9 × 10^7 cells; Brg1, 6.4 × 10^7 cells | native | cells resuspended in 2 mL low detergent buffer; digestions were performed with 15 U/mL MNase -> helicos; Chromatin from MEFs of all genotypes (~4 million cells/genotype) was also digested with MNase at two different concentrations (0.2 U or 4 U per 100-µL reaction volume)-> solexa | agarose gel (150bp) |

## Reported characteristics of nucleosome maps in mammals

### *TSS alignments vs. transcription*

The organization of nucleosomes at a promoter can have a regulatory effect on transcription. Nucleosomes can block transcription factors and polymerase from binding or can provide a substrate for protein complexes to bind and exert a regulatory effect. Thus, analysis of the nucleosome occupancy of the region surrounding the TSS can tell us about the physical and regulatory state of these chromatin regions. Many genome-wide MNase studies have reported differences in aggregation plot patterns at transcription start sites (TSSs) depending on transcription, including those done on mammalian cells[13,72,92,101,102,104,106]. Generally, occupancy differs around the TSS in genes with varying levels of transcription. Changes can be classified as changes in: the degree of occupancy (frequency of tags piling up across the TSS), the relative height of a particular nucleosome's occupancy--specifically the +1 or -1 nucleosomes, the spacing of the nucleosomes--specifically the phased nucleosomes emanating out from the +1 and -1 nucleosomes, or the relative extent of depletion at the NFR. Active promoters have relatively deep NFRs and well-defined phasing, and silent genes have a very shallow NFR with a single nucleosome detectable at the +1 location [106-111]. Changes in any of the above have been interpreted to be caused by changes in the transcription machinery, nucleosome modifications or subunit changes, or remodeling proteins or complexes.

Nucleosome spacing increases at the TSS +1 and -1 nucleosomes correlating with changes in transcription were also reported, but again, not all studies saw or reported such an effect[13,103]. More specifically, one study reports that nucleosome phasing correlates with RNA PolII positioning, with poised genes looking similar to active genes and stalled PolII contributing

to the inactive genes' nucleosome profiles. This same study suggests that these changes at the TSS are because of the loss of nucleosomes with specialized subunits (H2AZ) or specific modifications[103].

In yeast it has been shown that perturbations to cell state such as heat shock can alter the promoters of genes changing transcription. Human CD4+ cells activated by TCR signaling undergo a transcriptional response at a defined set of genes. When the promoters of these genes are assessed for nucleosome occupancy differences in bulk, minor changes are found. The -1 nucleosome appears unchanged and the +1 and +2 nucleosome levels increase. The authors of this study hypothesize that because the genes undergo a rapid transcriptional response they are already poised in a conformation that is conducive to transcription, and in fact they see little difference in the -1 nucleosomes of genes whose transcription changes and the genes that were already actively transcribed in the cells before stimulation[103].

### *Global occupancy*

Because chromatin remodeling has been shown to be widespread and crucial to transcriptional control it was somewhat surprising when studies in mammals showed that nucleosomes have more consistent global patterns of occupancy than would be expected by chance across cell types. In granulocytes and T cells, Valouev et al. saw positioning preferences in aggregate at regulatory elements but conclude the majority of the genome has little nucleosome positioning. They also comment on barrier model, defining one potential barrier as 'container sites', or a region of G/C flanked by A/T. Other potential barriers mentioned in this study are stalled polymerase and regulatory factors[13]. In a different set of cell types, seven human lymphoblastoid cell lines, Gaffney et al. performed paired-end MNase–seq. They found

that less than 10% of nucleosomes in the genome are moderately- to well-positioned, though there are also arrays of nucleosomes enriched at regulatory sites. It is suggested that these arrays may be formed by positioning nucleosomes against barriers as well, those barriers being proteins or protein complexes bound specifically to the DNA.[100]

### *Transcription factors and occupancy*

Transcription factors control the flow of genetic information in DNA to RNA. How they interact with chromatin to exert their regulatory effect is not fully understood. Transcription factors, beyond associating with arrays of phased nucleosomes, have been shown to rearrange nucleosomes *in vitro*. A well-studied transcription activator, Gal4, binds a nucleosome *in vitro* and forms a complex that is dissociated into nucleosomes or Gal4 bound to DNA when competitor DNA is added[112]. The binding of a protein to DNA and subsequent structural rearrangement of chromatin can clear the way for additional factors to bind, as is the case with the glucocorticoid receptor binding a nucleosome in the MMTV system[20,113-115]. This binding results in a change in DNaseI accessibility of the locus and NF1 factor binding. Another transcription factor that has been shown to rearrange nucleosomes *in vitro* is FoxA1. A pioneer transcription factor, FoxA (more specifically FoxA1 and FoxA2) can bind nucleosomal DNA and act as a 'pioneer' for more factors by rearranging nucleosomes and opening up the chromatin. This is sometimes called 'genetic potentiation'. To specifically address the question of the effect of Forkhead box factors on the nucleosome landscape surrounding those factors' binding sites Li et al. mapped the nucleosome occupancy of cells that were wild type and of cells that were lacking FoxA1 and FoxA2. While they could demonstrate that FoxA1 can bind nucleosomes through co-immunoprecipitation experiments, this study showed that FoxA1 and FoxA2

depletion does not impact nucleosome occupancy[72]. Furthermore, no clear difference in nucleosome occupancy was seen at genes whose regulation was altered by FoxA1 and FoxA2 knockout.

Intrinsic nucleosome forming capabilities refer to the ability of DNA sequences to be preferred binding sites for histone octamers. Several methods exist to determine the intrinsic nucleosome forming capabilities of a sequence, but many are based on determining enriched sequences in nucleosome occupancy data from different cell types. In yeast, regulatory elements such as promoters, enhancers, and TFBSs generally encode high intrinsic nucleosome occupancy, compared to human regulatory sequences which show the opposite effect.[116] It may be that in yeast some 'open' regulatory elements make sense in a single-celled organism and that in multicellular organisms nucleosomes protect more regulatory elements and must be moved for use of those sequences, however aspects of the methodology used may also play a role in this distinction, such as how a single MNase concentration profiles a smaller genome packaged into a nucleus. Further, *in vivo*, when factors do bind to DNA, phasing of nucleosomes around the bound locus can change[100]. One way to learn about how a bound factor exerts an influence physically is by looking at alignments of nucleosome occupancy averaged around the loci it binds. Many transcription factors recruit remodelers, and consequently may be changing chromatin. Further, a bound protein alone can change the nucleosome occupancy of a region by acting as a barrier. Kundaje et al. profiled nucleosome occupancy around more than 100 DNA binding proteins and assessed these binding sites for directional phasing. Interestingly, with the exception of the CTCF/cohesin complex which has a symmetric pattern around the binding site, asymmetry of nucleosome positioning around a

bound factor is predominant[102] Further, it was found that there are multiple phenotypes of nucleosome occupancy around the TSS. The authors clustered occupancy at TSSs by shape and found that some patterns of occupancy correlate well with high levels of expression and others with low expression. This bolsters the above point that expression state of a group of genes correlates with the pattern of nucleosome occupancy around their TSSs. Exactly what phasing around transcription factor binding sites says about the regulation of chromatin structure remains unclear, but it provides a starting point for future mechanistic studies.

Arrays of phased nucleosomes are a common feature seen in occupancy maps. These arrays tend to be seen at regulatory regions, most notably at the TSS, but also around transcription factors. To better classify chromatin structure Gaffney et al. used the chromatin states classified in Ernst et al.[117] as a comparison to nucleosome occupancy. These states were determined using maps of histone marks to define regions that had clusters of marks occurring together across the genome. This comparison showed an enrichment of nucleosome arrays at promoters, insulators, enhancers, and heterochromatin[100]. Seemingly, nucleosomes position next to where factors bind in the genome, with more than one nucleosome 'stacking up' against the factor, which is called barrier or statistical positioning. It is unclear what all the mechanisms are that produce this kind of occupancy pattern.

### *Effect of remodelers on nucleosome occupancy patterns*

With the first genome-wide nucleosome occupancy maps came the discovery of interesting chromatin organizational features. As mentioned above, little is known about the specific forces that create these features, such as the stereotypic patterns of occupancy around transcribed and silent TSSs, and the phased nucleosomes around some TF binding sites.

Remodeling proteins likely play a big role in the formation of these features. There are at least four remodeling complex families, SWI/SNF, ISWI, INO80/SWR1, and CHD. Proteins in these complexes have the ability to use ATP to move or destabilize nucleosomes; a well-studied example of one of these proteins is Brg1. Brg1 is an important part of the SWI/SNF family of remodeling complexes.

During differentiation of HSCs to erythrocytes, GATA1 binding sites, used as a proxy for enhancers, show a nucleosome shift away from the binding site. Brg1 co-binds these Gata1 enhancers, and when Brg1 is knocked down nucleosomes flanking the Gata1 site shift away from the binding site. However, Gata1 binding was found to be independent of Brg1 action. In cells with wild type Brg1, Tal1, a transcription factor, could bind efficiently to Gata1 sites; however in the knockdown this was not the case. Taken together the conclusion is that Brg1 shifts nucleosomes away from the Gata1 sites allowing TAL1 to bind[52]. Complicating the story is the fact that a subset of GATA1 sites were co-occupied with a nucleosome, potentially representing a separate class of nucleosome-bound GATA1. While this subset of GATA1 binding was not further investigated in this manuscript, one hypothesis is that some factors first bind a nucleosome and recruit remodeling machinery to alter the locus. Much work must be done to explore this hypothesis, which is prevalent in work done examining nucleosome occupancy in different cell states.

In mouse embryonic fibroblasts, the reduction of Brg1 alters the averaged TSS profile for a large number of diverse genes. The height of the +1 and -1 nucleosomes is greatly reduced on average in Brg1 depleted cells, and the phasing of nucleosomes is altered such that nucleosomes emanating out past the +1 and -1 nucleosomes become closer to each other when

Brg1 is depleted. Further, the degree of Brg1 occupancy at a promoter correlates with the degree of occupancy of the nucleosomes flanking the NFR. Accordingly, when Brg1 is reduced the degree of occupancy of these enriched nucleosomes is also reduced, with the greatest reduction correlating to the Brg1 sites[104].

As mentioned above, the degree of occupancy in a TSS alignment correlates with transcription of the genes in that alignment. Genes with more transcription have a higher average occupancy with a more pronounced NDR. Changes in nucleosome occupancy caused by knocking down remodelers do not seem to have an effect on transcription, however[104]. This was also recently shown to be the case for a remodeler in yeast[118].

### *Nucleosome occupancy relationship with other chromatin features*

A cell has many ways to change the stability of a nucleosome through modification, and many ways to move or alter a nucleosome's position in chromatin to affect gene expression. The nucleosome particle itself is dynamic, 'breathing' and moving to DNA that is most favorable for stability. Thus, determining where nucleosomes sit along the chromatin should tell us, independent of understanding all of these modifications and remodeling events, which DNA is exposed to nuclear factors that bind free DNA, which is occluded from these factors, and what DNA is only accessible to factors that can bind a nucleosome. A map of nucleosome occupancy provides the story of what the regulatory state of the chromatin in a cell is.

With whole genome sequencing researchers have been able to interrogate all classes of genomic elements for interesting nucleosome occupancy traits. In a system of T cell activation Schones et al. examined a set of conserved non-coding genomic regions that act like enhancers, and saw nucleosome depletion at these sequences after stimulation[106]. This finding is in

keeping with what has been traditionally assumed: that nucleosomes must be remodeled in a way to grant access to regulatory DNA in a cell type-specific manner. Additionally it was seen in this study that genes with poised polymerase were similar, in aggregate, to expressed genes in terms of nucleosome occupancy at the TSS. When polymerase is at the TSS a nucleosome-depleted region is seen in averaged TSS nucleosome occupancy profiles, along with phased nucleosomes emanating from the TSS. The averaged TSS plots for genes with no associated polymerase showed no NDR or phasing. Upon stimulus poised genes that became active had an aggregate shift towards less nucleosome occupancy downstream of the TSS, with no change upstream. Genes that are repressed upon stimulus have higher overall occupancy upstream and downstream of the TSS[106]. The TSS then has a characteristic architecture when engaged by the transcription machinery, regardless of transcription state, suggesting multiple conformations of occupancy are important for more than just 'on' and 'off' states of gene expression.

In addition to studying what happens at promoters and enhancers specifically, there has been a flurry of reports in the past few years that focus extensively on covalent histone tail modifications as a proxy for understanding the physical chromatin regulatory state. Regions marked with covalent histone tail modifications often correlate with evidence of a physical state of chromatin, and are usually called 'open' or 'closed'. The combination of histone mark ChIP-seq maps and MNase-seq maps has begun to be used to characterize the physical state of chromatin, although much more work remains to be done. Valouev et al. compared a variety of histone marks and determined the nucleosome spacing at enhancers (H3K4me1, H3K27ac, H3K36ac), and found that active promoter-associated domains had the shortest spacing

between nucleosomes[13]. A study in mouse ESCs found that H3K9me3 associated regions (typically correlated with repressed genes) are more occupied by nucleosomes than are sites associated with H3K27ac and H3K9ac (associated with active genes)[101]. This study also found that average nucleosome repeat length increased during differentiation by 5-7 base pairs[101].

Overall, mammalian nucleosome occupancy shows the same fundamental features as lower organisms, notably yeast. Not only do averaged nucleosome occupancy profiles show a characteristic TSS alignment profile, slight pattern changes are commensurate with transcriptional changes. Enhancers look open; arrays of nucleosomes occur around some bound factors. Regions of the genome with particular histone marks or binding proteins correlate to patterns of nucleosome occupancy in a predicable way, supporting descriptions of 'open' and 'closed' chromatin. Globally, transcriptional activity and nucleosome occupancy profiles have correlated features, however when comparing nucleosome occupancy in different cell lines the differences in transcription do not significantly correlate with changes in occupancy. Additionally, direct comparison of occupancy between cell states shows that the vast majority of the genome has very similar nucleosome occupancy[13,72,103].

### Open questions

Many of the findings of mammalian nucleosome occupancy studies are based on averaged effects at known regulatory regions. Because access to the DNA is so important and highly regulated it seems like we should see a multitude of effects of regulation in MNase-seq nucleosome occupancy maps, especially at the local level. In fact, yeast studies show very reproducible TSS alignments across studies and strikingly clearly positioned nucleosomes at local regions of DNA. Yeast studies often show changes at single nucleosome loci in different

conditions, displaying a clear effect of regulatory machinery upon stimulus. We rarely see such effect in mammalian studies. Although there is an obvious distinction between mammals and yeast in genome size and complexity, there may be more to the explanation of why mammalian nucleosome occupancy features (such as TSS alignments) don't look as uniform across publications as does yeast; it will be interesting to see if the differences can be explained by biological differences between the species.

A look at the methodology used across mammal MNase mapping experiments suggests that different populations of DNA fragments are investigated in different experiments. Preparation of material for MNase digests is important. Across mammalian MNase-seq publications, initial cell preparation ranges from freezing whole cell populations or tissue in liquid nitrogen and crushing, to hypotonic buffer-mediated lysis, to digestion in native conditions, to cross-linking, with or without salt extraction. Each of these protocols presumably affects the proteins bound to chromatin in different ways, with proteins bound differently depending on salts, length of processing, and extent of crosslinking.

Even more simply the population of cells being assessed is important for any study of chromatin—a tissue sample will have several cell types with several global chromatin conditions within. Any chromatin preparation made from a population of cells (heterogeneous or not) likely has a heterogeneous mix of marks or architecture at any given locus, so cell number should a consideration, especially if samples are to be compared. Along these lines genome size will determine the amount of sequence that must be obtained to achieve an appropriate amount of coverage of reads at any given locus. It is likely that sub-optimal sequence coverage is a major reason we don't see well-positioned nucleosomes in mammalian

occupancy studies. Temperature and time of digestion vary greatly across MNase-mapping experiments in the literature with no standard for how to asses the result of a digestion (when noted in the methods the choice of sample conditions can be because of the appearance of a DNA ladder on a gel or proportion of DNA in a band or bands, although even these conditions vary greatly). Minimal attention has been given in the literature to the fact that different levels of digestion with a nuclease will liberate different populations of fragments[91-93,119]; with most MNase-seq experiments using the conventional single digestion condition and sometimes no explanation of that condition and how it was chosen or the appearance of the resulting MNase ladder in the manuscript. Simply stated, there is room for improvement on techniques using MNase to assess nucleosome occupancy.

## Pluripotent versus differentiated cells

The regulation of chromatin structure is important for the state of a cell, and this is particularly well studied in the pluripotent cell state. Pluripotent chromatin is said to be 'open' compared to lineage-committed cells[120-128]. Chromatin openness entails an assumption of accessibility of DNA and infers a non-compacted structure. This often-stated characterization is based on a set of mostly cytological and biochemical data. Interestingly, much of the evidence describing the open state of chromatin in ESCs has shown a similar state in cells reprogrammed to the pluripotent state from differentiated cells (iPSC)[126,127,129-131]. There are, however, ample studies suggesting that remnants of the epigenetic state of reprogrammed cells remain in iPSCs[132-136] and these cell of origin signatures could have a huge impact on the use of these reprogrammed cells in therapeutic and research contexts.

The most basic line of evidence that suggests ESCs have an open chromatin state is transcription itself. Radiolabeling of new RNAs in ESCs and differentiated NPCs showed a 2 fold increase in overall transcription in ESC, with regions of the genome typically repressed in differentiated cells, such as satellite, LINE, and SINE repeats, being transcribed in ESCs[128]. Experiments using MNase to assess the amount of histones released over time showed that nearly all histones in pluripotent cell were released within 10 minutes, while in differentiated cells it took more than twice as long to liberate most histones under the same conditions[80].

Differences in the compaction state of chromatin DNA in the context of the mammalian nucleus can be seen using electron microscopy, with tightly compacted chromatin appearing dark (heterochromatin) and less compacted chromatin appearing light. In mouse ES cells chromatin looks homogenous when observed using electron microscopy, and becomes more heterogeneous with sub-regions becoming more condensed as cells differentiate[128]. Additional evidence that chromatin is organized uniquely in pluripotent cells versus differentiated cells was shown through FISH by probing alpha satellite sequences, which are enriched in heterochromatic regions. This satellite-rich heterochromatin was more diffusely stained in pluripotent cells, and upon differentiation to neural progenitor cells the signal was more punctate and defined[80].

In addition to using FISH to visualize heterochromatin under the microscope, histone variants and covalent histone modifications can be immuno-stained and viewed on a 'macro' scale as well. Upon differentiation the localization of heterochromatin protein HP1alpha and repressive mark H3K9me3 change drastically from being spread diffusely in the nucleus in pluripotent cells to being organized in punctate foci in neural progenitor cells[80]. Overall the size

of these foci becomes smaller and their number increases with differentiation; these changes suggest a mechanism of chromatin condensation impacting the chromatin to set cell fate. Supporting this model, acetylation of H3 and H4, typically marks of active chromatin, are reduced in differentiation[80].

The above studies show that the location of certain proteins and modifications to proteins in the cell changes in a way that suggests chromatin is being locally condensed or opened in cell fate specification. The correlation of this phenomena with gene repression and activation suggests these loci are being locked down or opened up to set cell state. These studies don't say anything about protein turnover at these locations, however. A measure of protein dynamics in the cell is Fluorescence Recovery After Photobleaching (FRAP). FRAP experiments in pluripotent and differentiated cells showed HP1alpha has different dynamics in the heterochromatin of these two cell types, with a faster turnover in the pluripotent cells. Similarly, H2B and H3 fluorescence-tagged histones turned over faster in pluripotent cells than in differentiated cells. The recovery after photo bleaching of all of the above histone proteins was characterized by a rapid initial recovery that could indicate a loosely bound or soluble pool of histone proteins available in the pluripotent cells[80].

As noted previously, many studies aiming to characterize the chromatin state of a cell profile histone tail modifications. In the pluripotent to differentiated transition histone marks change concurrent with silencing[137-139]. A pairing of H3K4me3 and K3K27me3 not seen in other cell types occurs at poised genes in pluripotent cells, termed bivalent[50,140]. A collection of genome-wide maps of chromatin marks lends support to the characterization of pluripotent chromatin, in both embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSC), being

globally 'open'[141-143]. Of note, many of the known chromatin characteristics of ESCs have been assessed in iPSC and are reset during reprogramming[79,125.].

The processes that maintain pluripotency or promote differentiation rely on multiple chromatin regulators. These regulators include a chromatin-remodeling complex unique to ESCs.[120-122,144-146] It is also known that the highly studied pluripotency factors Oct4, Sox2, and Nanog recruit chromatin remodeling factors to exert their effect in reprogramming[147]. All of the above taken together suggest that the chromatin state in pluripotent cells is more 'open' than in differentiated cell types, although direct physical evidence of this is incomplete.

## Conclusion

Regulatory factors require access to DNA, and chromatin state impacts this access. A variety of modifications to the histone octamer in nucleosomes as well as chemical modifications to these histones and DNA can alter the stability of nucleosomes. Further, these modifications can act as signals to proteins with binding domains specific to the modification that exert an effect on chromatin state. Remodelers are one class of protein that can recognize these marks and change the physical state of chromatin, making it more open or closed to regulatory factors.

Direct chemical modification and remodeling of nucleosomes only partly explain chromatin's dynamic state. Chromatin, whose basic component is the nucleosome, is also fundamentally dynamic. Nucleosome DNA "breathes" by unwrapping and rewrapping, and nucleosomes can slide along the DNA. In fact, nucleosomes *in vitro* prefer certain sequences. It follows that nucleosomes located in non-preferred locations suggest a regulatory force at

action. Thus, a map of nucleosomes along chromatin can show us both the impact of regulatory forces and the accessibility of DNA to regulatory factors.

MNase has been used to probe the accessibility of chromatin and to map nucleosome occupancy for decades. Recent advances in sequencing technology have allowed genome wide studies of nucleosome occupancy. Mammalian nucleosome occupancy maps show several characteristic features: a stereotypic averaged nucleosome occupancy profile around the transcription start site including a nucleosome depleted region that varies in depth depending on transcription level, phasing of nucleosomes around some DNA binding proteins like transcription factors and NDRs correlating with DNaseI sensitive sites correlating with regulatory elements, suggesting an open state of these regions.

Nucleosome occupancy maps have also shown a peculiar feature—there are often very few differences in maps between cell types, even when large-scale chromatin structural differences are known to occur. In Chapter 2 below I map nucleosome occupancy in two cell types known to have very different chromatin states with careful consideration of experimental conditions and data analysis to begin to address this feature of MNase seq maps. It is also unclear why there are discrepancies between studies of single loci and genome wide maps, for example in the case of FoxA1/FoxA2 regulation. A possible explanation is that current methodology does not profile the accessibility of all physical states of chromatin, as only one or two MNase digestion conditions are used in published studies in mammalian cells. Because chromatin organization is such a highly regulated and important factor in cell function, deeply understanding how MNase can be used to profile chromatin accessibility is of interest. To address this question I performed experiments and analysis that expanded on traditional

MNase methodology. These studies provide insight into the action of a long-used tool and the state of chromatin accessibility.

## References


1 Petesch, S. & Lis, J. Overcoming the nucleosome barrier during transcript elongation. *Trends in genetics : TIG* **28**, 285-294, doi:10.1016/j.tig.2012.02.005 (2012).

2 Steinman, R. & Moberg, C. A triple tribute to the experiment that transformed biology. *The Journal of experimental medicine* **179**, 379-384 (1994).

3 Olins, D. & Olins, A. Chromatin history: our view from the bridge. *Nature reviews. Molecular cell biology* **4**, 809-814, doi:10.1038/nrm1225 (2003).

4 Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128**, 707-719 (2007).

5 Morse, R. H. Transcription factor access to promoter elements. *J Cell Biochem* **102**, 560-570 (2007).

6 Luger, K., Mäder, A., Richmond, R., Sargent, D. & Richmond, T. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-260, doi:10.1038/38444 (1997).

7 Kornberg, R. Chromatin structure: a repeating unit of histones and DNA. *Science (New York, N.Y.)* **184**, 868-871, doi:10.1126/science.184.4139.868 (1974).

8 Worcel, A., Han, S. & Wong, M. Assembly of newly replicated chromatin. *Cell* **15**, 969-977 (1978).

9 Verreault, A. De novo nucleosome assembly: new pieces in an old puzzle. *Genes & development* **14**, 1430-1438, doi:10.1101/gad.14.12.1430 (2000).

10 Clark, R. & Felsenfeld, G. Structure of chromatin. *Nature: New biology* **229**, 101-106, doi:10.1038/newbio229101a0 (1971).

11 Itzhaki, R. Studies on the accessibility of deoxyribonucleic acid in deoxyribonucleoprotein to cationic molecules. *The Biochemical journal* **122**, 583-592 (1971).

12 Mirsky, A. The structure of chromatin. *Proceedings of the National Academy of Sciences of the United States of America* **68**, 2945-2948, doi:10.1073/pnas.68.12.2945 (1971).

13 Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516-520, doi:10.1038/nature10002 (2011).

14 Demeret, C., Vassetzky, Y. & Méchali, M. Chromatin remodelling and DNA replication: from nucleosomes to loop domains. *Oncogene* **20**, 3086-3093, doi:10.1038/sj.onc.1204333 (2001).

15 Hara, R., Mo, J. & Sancar, A. DNA damage in the nucleosome core is refractory to repair by human excision nuclease. *Molecular and cellular biology* **20**, 9173-9181 (2000).

16 Getun, I. V., Wu, Z. K., Khalil, A. M. & Bois, P. R. Nucleosome occupancy landscape and dynamics at mouse recombination hotspots. *EMBO reports* **11**, 555-560, doi:10.1038/embor.2010.79 (2010).

17 Alexiadis, V. & Kadonaga, J. T. Strand pairing by Rad54 and Rad51 is enhanced by chromatin. *Genes & development* **16**, 2767-2771, doi:10.1101/gad.1032102 (2002).

18 Almer, A., Rudolph, H., Hinnen, A. & Hörz, W. Removal of positioned nucleosomes from the yeast PHO5 promoter upon PHO5 induction releases additional upstream activating DNA elements. *The EMBO journal* **5**, 2689-2696 (1986).

19 Owen-Hughes, T. & Workman, J. L. Remodeling the chromatin structure of a nucleosome array by transcription factor-targeted trans-displacement of histones. *The EMBO journal* (1996).

20 Richard-Foy, H. & Hager, G. Sequence-specific positioning of nucleosomes over the steroid-inducible MMTV promoter. *The EMBO journal* **6**, 2321-2328 (1987).

21 Takahata, S., Yu, Y. & Stillman, D. J. FACT and Asf1 Regulate Nucleosome Dynamics and Coactivator Binding at the< i> HO</i> Promoter. *Molecular cell*, doi:10.1016/j.molcel.2009.04.010 (2009).

22 Lorch, Y., LaPointe, J. & Kornberg, R. Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell* **49**, 203-210, doi:10.1016/0092-8674(87)90561-7 (1987).

23 Satchwell, S., Drew, H. & Travers, A. Sequence periodicities in chicken nucleosome core DNA. *Journal of molecular biology* **191**, 659-675, doi:10.1016/0022-2836(86)90452-3 (1986).

24 Lowary, P. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of molecular biology* **276**, 19-42, doi:10.1006/jmbi.1997.1494 (1998).

25 Widom, J. Short-range order in two eukaryotic genomes: relation to chromosome structure. *Journal of molecular biology* **259**, 579-588, doi:10.1006/jmbi.1996.0341 (1996).

26 Tolstorukov, M., Colasanti, A., McCandlish, D., Olson, W. & Zhurkin, V. A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *Journal of molecular biology* **371**, 725-738, doi:10.1016/j.jmb.2007.05.048 (2007).

27 Tolstorukov, M., Choudhary, V., Olson, W., Zhurkin, V. & Park, P. nuScore: a web-interface for nucleosome positioning predictions. *Bioinformatics (Oxford, England)* **24**, 1456-1458, doi:10.1093/bioinformatics/btn212 (2008).

28 Reynolds, S., Bilmes, J. & Noble, W. Learning a weighted sequence model of the nucleosome core and linker yields more accurate predictions in Saccharomyces cerevisiae and Homo sapiens. *PLoS computational biology* **6**, doi:10.1371/journal.pcbi.1000834 (2010).

29 Weber, C. & Henikoff, S. Histone variants: dynamic punctuation in transcription. *Genes & development* **28**, 672-682, doi:10.1101/gad.238873.114 (2014).

30 Howman, E. *et al.* Early disruption of centromeric chromatin organization in centromere protein A (Cenpa) null mice. *Proceedings of the National Academy of Sciences of the United States of America* **97**, 1148-1153, doi:10.1073/pnas.97.3.1148 (2000).

31 Tachiwana, H. *et al.* Structural basis of instability of the nucleosome containing a testis-specific histone variant, human H3T. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 10454-10459, doi:10.1073/pnas.1003064107 (2010).

32 Costanzi, C. & Pehrson, J. Histone macroH2A1 is concentrated in the inactive X chromosome of female mammals. *Nature* **393**, 599-601, doi:10.1038/31275 (1998).

33 Tolstorukov, M. *et al.* Histone variant H2A.Bbd is associated with active transcription and mRNA processing in human cells. *Molecular cell* **47**, 596-607, doi:10.1016/j.molcel.2012.06.011 (2012).

34 Barrero, M. *et al.* Macrohistone variants preserve cell identity by preventing the gain of H3K4me2 during reprogramming to pluripotency. *Cell reports* **3**, 1005-1011, doi:10.1016/j.celrep.2013.02.029 (2013).

35 Nekrasov, M. *et al.* Histone H2A.Z inheritance during the cell cycle and its impact on promoter organization and dynamics. *Nature structural & molecular biology* **19**, 1076-1083, doi:10.1038/nsmb.2424 (2012).

36 Obri, A. *et al.* ANP32E is a histone chaperone that removes H2A.Z from chromatin. *Nature* **505**, 648-653, doi:10.1038/nature12922 (2014).

37 Jin, C. *et al.* H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nature genetics* **41**, 941-945, doi:10.1038/ng.409 (2009).

38 Tolstorukov, M., Kharchenko, P., Goldman, J., Kingston, R. & Park, P. Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. *Genome research* **19**, 967-977, doi:10.1101/gr.084830.108 (2009).

39 Skene, P. & Henikoff, S. Histone variants in pluripotency and disease. *Development (Cambridge, England)* **140**, 2513-2524, doi:10.1242/dev.091439 (2013).

40 Graber, M., Schweinfest, C., Reed, C., Papas, T. & Baron, P. Isolation of differentially expressed genes in carcinoma of the esophagus. *Annals of surgical oncology* **3**, 192-197, doi:10.1007/BF02305800 (1996).

41 Maze, I., Noh, K.-M., Soshnev, A. & Allis, C. Every amino acid matters: essential contributions of histone variants to mammalian development and disease. *Nature reviews. Genetics*, doi:10.1038/nrg3673 (2014).

42 Fan, Y., Sirotkin, A., Russell, R., Ayala, J. & Skoultchi, A. Individual somatic H1 subtypes are dispensable for mouse development even in mice lacking the H1(0) replacement subtype. *Molecular and cellular biology* **21**, 7933-7943, doi:10.1128/MCB.21.23.7933-7943.2001 (2001).

43 Izzo, A. *et al.* The genomic landscape of the somatic linker histone subtypes H1.1 to H1.5 in human cells. *Cell reports* **3**, 2142-2154, doi:10.1016/j.celrep.2013.05.003 (2013).

44 Kouzarides, T. Chromatin modifications and their function. *Cell* **128**, 693-705, doi:10.1016/j.cell.2007.02.005 (2007).

45 Strahl, B. & Allis, C. The language of covalent histone modifications. *Nature* **403**, 41-45, doi:10.1038/47412 (2000).

46 Jenuwein, T. & Allis, C. Translating the histone code. *Science (New York, N.Y.)* **293**, 1074-1080, doi:10.1126/science.1063127 (2001).

47 Craig, L. P. & Marc-André, L. Histones and histone modifications. *Current Biology*, doi:10.1016/j.cub.2004.07.007 (2004).

48 Patrick, A. G. A tale of histone modifications. *Genome Biology*, doi:10.1186/gb-2001-2-4-reviews0003 (2001).

49 Zhang, H., Gao, L., Anandhakumar, J. & Gross, D. Uncoupling transcription from covalent histone modification. *PLoS genetics* **10**, doi:10.1371/journal.pgen.1004202 (2014).

50 Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125**, 315-326 (2006).

51 Shen, W. *et al.* Solution structure of human Brg1 bromodomain and its specific binding to acetylated histone tails. *Biochemistry* **46**, 2100-2110, doi:10.1021/bi0611208 (2007).

52 Hu, G. *et al.* Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome research* **21**, 1650-1658, doi:10.1101/gr.121145.111 (2011).

53 Saha, A., Wittmeyer, J. & Cairns, B. R. Chromatin remodelling: the industrial revolution of DNA around histones. *Nature reviews. Molecular cell biology* **7**, 437-447, doi:10.1038/nrm1945 (2006).

54 Schwartzentruber, J. *et al.* Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature* **482**, 226-231, doi:10.1038/nature10833 (2012).

55 Wu, G. *et al.* Somatic histone H3 alterations in pediatric diffuse intrinsic pontine gliomas and non-brainstem glioblastomas. *Nature genetics* **44**, 251-253, doi:10.1038/ng.1102 (2012).

56 Khuong-Quang, D.-A. *et al.* K27M mutation in histone H3.3 defines clinically and biologically distinct subgroups of pediatric diffuse intrinsic pontine gliomas. *Acta neuropathologica* **124**, 439-447, doi:10.1007/s00401-012-0998-0 (2012).

57 Lewis, P. *et al.* Inhibition of PRC2 activity by a gain-of-function H3 mutation found in pediatric glioblastoma. *Science (New York, N.Y.)* **340**, 857-861, doi:10.1126/science.1232245 (2013).

58 Chan, K.-M. *et al.* The histone H3.3K27M mutation in pediatric glioma reprograms H3K27 methylation and gene expression. *Genes & development* **27**, 985-990, doi:10.1101/gad.217778.113 (2013).

59 Wagner, E. & Carpenter, P. Understanding the language of Lys36 methylation at histone H3. *Nature reviews. Molecular cell biology* **13**, 115-126, doi:10.1038/nrm3274 (2012).

60 Margueron, R. & Reinberg, D. The Polycomb complex PRC2 and its mark in life. *Nature* **469**, 343-349, doi:10.1038/nature09784 (2011).

61 Mersfelder, E. & Parthun, M. The tale beyond the tail: histone core domain modifications and the regulation of chromatin structure. *Nucleic acids research* **34**, 2653-2662, doi:10.1093/nar/gkl338 (2006).

62 Freitas, M., Sklenar, A. & Parthun, M. Application of mass spectrometry to the identification and quantification of histone post-translational modifications. *Journal of cellular biochemistry* **92**, 691-700, doi:10.1002/jcb.20106 (2004).

63 Cosgrove, M., Boeke, J. & Wolberger, C. Regulated nucleosome mobility and the histone code. *Nature structural & molecular biology* **11**, 1037-1043, doi:10.1038/nsmb851 (2004).

64 Mersfelder, E. L. & Parthun, M. R. The tale beyond the tail: histone core domain modifications and the regulation of chromatin structure. *Nucleic acids research* **34**, 2653-2662, doi:10.1093/nar/gkl338 (2005).

65 Weber, C., Ramachandran, S. & Henikoff, S. Nucleosomes Are Context-Specific, H2A.Z-Modulated Barriers to RNA Polymerase. *Molecular cell* **53**, 819-830, doi:10.1016/j.molcel.2014.02.014 (2014).

66 Lidor Nili, E. *et al.* p53 binds preferentially to genomic regions with high DNA-encoded nucleosome occupancy. *Genome research* **20**, 1361-1368, doi:10.1101/gr.103945.109 (2010).

67 Sahu, G. *et al.* p53 binding to nucleosomal DNA depends on the rotational positioning of DNA response element. *The Journal of biological chemistry* **285**, 1321-1332, doi:10.1074/jbc.M109.081182 (2010).

68 Ballaré, C. *et al.* Nucleosome-driven transcription factor binding and gene regulation. *Molecular cell* **49**, 67-79, doi:10.1016/j.molcel.2012.10.019 (2013).

69 Laptenko, O., Beckerman, R., Freulich, E. & Prives, C. p53 binding to nucleosomes within the p21 promoter in vivo leads to nucleosome loss and transcriptional activation. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 10385-10390, doi:10.1073/pnas.1105680108 (2011).

70 Chaya, D., Hayamizu, T., Bustin, M. & Zaret, K. Transcription factor FoxA (HNF3) on a nucleosome at an enhancer complex in liver chromatin. *The Journal of biological chemistry* **276**, 44385-44389, doi:10.1074/jbc.M108214200 (2001).

71 Cirillo, L. A. & Zaret, K. S. An early developmental transcription factor complex that is more stable on nucleosome core particles than on free DNA. *Molecular cell* **4**, 961-969 (1999).

72 Li, Z., Schug, J., Tuteja, G., White, P. & Kaestner, K. The nucleosome map of the mammalian liver. *Nature structural & molecular biology* **18**, 742-746, doi:10.1038/nsmb.2060 (2011).

73 Li, Z. *et al.* Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell* **151**, 1608-1616, doi:10.1016/j.cell.2012.11.018 (2012).

74 Anderson, J. D. & Widom, J. Sequence and position-dependence of the equilibrium accessibility of nucleosomal DNA target sites. *Journal of molecular biology* **296**, 979-987, doi:10.1006/jmbi.2000.3531 (2000).

75 Li, G., Levitus, M., Bustamante, C. & Widom, J. Rapid spontaneous accessibility of nucleosomal DNA. *Nature structural & molecular biology* **12**, 46-53, doi:10.1038/nsmb869 (2005).

76 Kraushaar, D. *et al.* Genome-wide incorporation dynamics reveal distinct categories of turnover for the histone variant H3.3. *Genome biology* **14**, doi:10.1186/gb-2013-14-10-r121 (2013).

77 Sudarsanam, P. & Winston, F. The Swi/Snf family nucleosome-remodeling complexes and transcriptional control. *Trends in genetics : TIG* **16**, 345-351 (2000).

78 Yaniv, M. Chromatin remodeling: from transcription to cancer. *Cancer genetics*, doi:10.1016/j.cancergen.2014.03.006 (2014).

79 Bouazoune, K. & Kingston, R. E. Chromatin remodeling by the CHD7 protein is impaired by mutations that cause human developmental disorders. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 19238-19243, doi:10.1073/pnas.1213825109 (2012).

80 Meshorer, E. *et al.* Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Developmental cell* **10**, 105-116, doi:10.1016/j.devcel.2005.10.017 (2006).

81 Giresi, P. G. & Lieb, J. D. Isolation of active regulatory elements from eukaryotic chromatin using FAIRE (Formaldehyde Assisted Isolation of Regulatory Elements). *Methods (San Diego, Calif.)* **48**, 233-239, doi:10.1016/j.ymeth.2009.03.003 (2009).

82 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213-1218, doi:10.1038/nmeth.2688 (2013).

83 Allan, J., Fraser, R., Owen-Hughes, T. & Keszenman-Pereyra, D. Micrococcal nuclease does not substantially bias nucleosome mapping. *Journal of molecular biology* **417**, 152-164, doi:10.1016/j.jmb.2012.01.043 (2012).

84 Heins, J., Suriano, J., Taniuchi, H. & Anfinsen, C. Characterization of a nuclease produced by Staphylococcus aureus. *The Journal of biological chemistry* **242**, 1016-1020 (1967).

85 Hewish, D. R. & Burgoyne, L. A. Chromatin sub-structure. The digestion of chromatin DNA at regularly spaced sites by a nuclear deoxyribonuclease. *Biochemical and biophysical research communications* **52**, 504-510 (1973).

86 Dingwall, C., Lomonossoff, G. & Laskey, R. High sequence specificity of micrococcal nuclease. *Nucleic acids research* **9**, 2659-2673, doi:10.1093/nar/9.12.2659 (1981).

87 Hörz, W. & Altenburger, W. Sequence specific cleavage of DNA by micrococcal nuclease. *Nucleic acids research* **9**, 2643-2658, doi:10.1093/nar/9.12.2643 (1981).

88 Locke, G., Tolkunov, D., Moqtaderi, Z., Struhl, K. & Morozov, A. High-throughput sequencing reveals a simple model of nucleosome energetics. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 20998-21003, doi:10.1073/pnas.1003838107 (2010).

89 Chung, H.-R. *et al.* The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS One* **5**, doi:10.1371/journal.pone.0015754 (2010).

90 Fan, X. *et al.* Nucleosome depletion at yeast terminators is not intrinsic and can occur by a transcriptional mechanism linked to 3'-end formation. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 17945-17950, doi:10.1073/pnas.1012674107 (2010).

91 Rizzo, J., Bard, J. & Buck, M. Standardized collection of MNase-seq experiments enables unbiased dataset comparisons. *BMC molecular biology* **13**, 15, doi:10.1186/1471-2199-13-15 (2012).

92 Weiner, A., Hughes, A., Yassour, M., Rando, O. & Friedman, N. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome research* **20**, 90-100, doi:10.1101/gr.098509.109 (2010).

93 Xi, Y., Yao, J., Chen, R., Li, W. & He, X. Nucleosome fragility reveals novel functional states of chromatin and poises genes for activation. *Genome research* **21**, 718-724, doi:10.1101/gr.117101.110 (2011).

94 Teves, S. S. & Henikoff, S. Salt fractionation of nucleosomes for genome-wide profiling. *Methods in molecular biology (Clifton, N.J.)* **833**, 421-432, doi:10.1007/978-1-61779-477-3_25 (2011).

95 Lee, W. *et al.* A high-resolution atlas of nucleosome occupancy in yeast. *Nature genetics* **39**, 1235-1244, doi:10.1038/ng2117 (2007).

96 Kornberg, R. & Stryer, L. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic acids research* **16**, 6677-6690, doi:10.1093/nar/16.14.6677 (1988).

97 Mavrich, T. *et al.* A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome research* **18**, 1073-1083, doi:10.1101/gr.078261.108 (2008).

98 Zhang, Z. *et al.* A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. *Science (New York, N.Y.)* **332**, 977-980, doi:10.1126/science.1200508 (2011).

99 Shivaswamy, S. *et al.* Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS biology* **6**, doi:10.1371/journal.pbio.0060065 (2008).

100 Gaffney, D. *et al.* Controls of nucleosome positioning in the human genome. *PLoS genetics* **8**, doi:10.1371/journal.pgen.1003036 (2012).

101 Teif, V. *et al.* Genome-wide nucleosome positioning during embryonic stem cell development. *Nature structural & molecular biology* **19**, 1185-1192, doi:10.1038/nsmb.2419 (2012).

102 Kundaje, A. *et al.* Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome research* **22**, 1735-1747, doi:10.1101/gr.136366.111 (2012).

103 Schones, D. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887-898, doi:10.1016/j.cell.2008.02.022 (2008).

104 Tolstorukov, M. *et al.* Swi/Snf chromatin remodeling/tumor suppressor complex establishes nucleosome occupancy at target promoters. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 10165-10170, doi:10.1073/pnas.1302209110 (2013).

105 Barozzi, I. *et al.* Coregulation of Transcription Factor Binding and Nucleosome Occupancy through DNA Features of Mammalian Enhancers. *Molecular cell*, doi:10.1016/j.molcel.2014.04.006 (2014).

106 Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887-898 (2008).

107 Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516-520, doi:10.1038/nature10002 (2011).

108 Teif, V. B. *et al.* Genome-wide nucleosome positioning during embryonic stem cell development. *Nature structural & molecular biology* **19**, 1185-1192, doi:10.1038/nsmb.2419 (2012).

109 Gaffney, D. J. *et al.* Controls of nucleosome positioning in the human genome. *PLoS genetics* **8**, e1003036, doi:10.1371/journal.pgen.1003036 (2012).

110 Li, Z., Schug, J., Tuteja, G., White, P. & Kaestner, K. H. The nucleosome map of the mammalian liver. *Nature structural & molecular biology* **18**, 742-746, doi:10.1038/nsmb.2060 (2011).

111 Tolstorukov, M. Y. *et al.* Swi/Snf chromatin remodeling/tumor suppressor complex establishes nucleosome occupancy at target promoters. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 10165-10170, doi:10.1073/pnas.1302209110 (2013).

112 Workman, J. & Kingston, R. Nucleosome core displacement in vitro via a metastable transcription factor-nucleosome complex. *Science (New York, N.Y.)* **258**, 1780-1784, doi:10.1126/science.1465613 (1992).

113 Bresnick, E., Rories, C. & Hager, G. Evidence that nucleosomes on the mouse mammary tumor virus promoter adopt specific translational positions. *Nucleic acids research* **20**, 865-870, doi:10.1093/nar/20.4.865 (1992).

114 Cordingley, M., Riegel, A. & Hager, G. Steroid-dependent interaction of transcription factors with the inducible promoter of mouse mammary tumor virus in vivo. *Cell* **48**, 261-270 (1987).

115 Zaret, K. & Yamamoto, K. Reversible and persistent changes in chromatin structure accompany activation of a glucocorticoid-dependent enhancer element. *Cell* **38**, 29-38, doi:10.1016/0092-8674(84)90523-3 (1984).

116 Tillo, D. *et al.* High nucleosome occupancy is encoded at human regulatory sequences. *PloS one* **5**, doi:10.1371/journal.pone.0009129 (2010).

117 Ernst, J. *et al.* Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature* **473**, 43-49, doi:10.1038/nature09906 (2011).

118 Gkikopoulos, T. *et al.* A role for Snf2-related nucleosome-spacing enzymes in genome-wide nucleosome organization. *Science (New York, N.Y.)* **333**, 1758-1760, doi:10.1126/science.1206097 (2011).

119 Bryant, G. O. *et al.* Activator control of nucleosome occupancy in activation and repression of transcription. *PLoS biology* **6**, 2928-2939, doi:10.1371/journal.pbio.0060317 (2008).

120 Gao, X. *et al.* ES cell pluripotency and germ-layer formation require the SWI/SNF chromatin remodeling component BAF250a. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 6656-6661 (2008).

121 Ho, L. *et al.* An embryonic stem cell chromatin remodeling complex, esBAF, is essential for embryonic stem cell self-renewal and pluripotency. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 5181-5186 (2009).

122 Loh, Y. H., Zhang, W., Chen, X., George, J. & Ng, H. H. Jmjd1a and Jmjd2c histone H3 Lys 9 demethylases regulate self-renewal in embryonic stem cells. *Genes & development* **21**, 2545-2557 (2007).

123 Meshorer, E. *et al.* Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Dev Cell* **10**, 105-116 (2006).

124 Fussner, E. *et al.* Constitutive heterochromatin reorganization during somatic cell reprogramming. *The EMBO journal* **30**, 1778-1789, doi:10.1038/emboj.2011.96 (2011).

125 Gaspar-Maia, A., Alajem, A., Meshorer, E. & Ramalho-Santos, M. Open chromatin in pluripotency and reprogramming. *Nature reviews. Molecular cell biology* **12**, 36-47, doi:10.1038/nrm3036 (2011).

126 Biran, A. & Meshorer, E. Concise review: chromatin and genome organization in reprogramming. *Stem cells (Dayton, Ohio)* **30**, 1793-1799, doi:10.1002/stem.1169 (2012).

127 Mattout, A., Biran, A. & Meshorer, E. Global epigenetic changes during somatic cell reprogramming to iPS cells. *Journal of molecular cell biology* **3**, 341-350, doi:10.1093/jmcb/mjr028 (2011).

128 Efroni, S. *et al.* Global transcription in pluripotent embryonic stem cells. *Cell stem cell* **2**, 437-447, doi:10.1016/j.stem.2008.03.021 (2008).

129 Fussner, E. *et al.* Constitutive heterochromatin reorganization during somatic cell reprogramming. *The EMBO journal* **30**, 1778-1789, doi:10.1038/emboj.2011.96 (2011).

130 Park, I.-H. *et al.* Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451**, 141-146, doi:10.1038/nature06534 (2008).

131 Kim, K. *et al.* Epigenetic memory in induced pluripotent stem cells. *Nature* (2010).

132 Lister, R. *et al.* Hotspots of aberrant epigenomic reprogramming in human induced pluripotent stem cells. *Nature* **471**, 68-73, doi:10.1038/nature09798 (2011).

133 Ruiz, S. *et al.* Identification of a specific reprogramming-associated epigenetic signature in human induced pluripotent stem cells. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 16196-16201, doi:10.1073/pnas.1202352109 (2012).

134 Chin, M. H. *et al.* Induced pluripotent stem cells and embryonic stem cells are distinguished by gene expression signatures. *Cell Stem Cell* **5**, 111-123 (2009).

135 Ghosh, Z. *et al.* Persistent donor cell gene expression among human induced pluripotent stem cells contributes to differences with human embryonic stem cells. *PLoS One* **5**, e8975 (2010).

136 Doi, A. *et al.* Differential methylation of tissue- and cancer-specific CpG island shores distinguishes human induced pluripotent stem cells, embryonic stem cells and fibroblasts. *Nat Genet* **41**, 1350-1353 (2009).

137 Lee, D., Hayes, J., Pruss, D. & Wolffe, A. A positive role for histone acetylation in transcription factor access to nucleosomal DNA. *Cell* **72**, 73-84 (1993).

138 Wolffe, A. Packaging principle: how DNA methylation and histone acetylation control the transcriptional activity of chromatin. *The Journal of experimental zoology* **282**, 239-244, doi:10.1002/(SICI)1097-010X(199809/10)282:1/2<239::AID-JEZ25>3.0.CO;2-N (1998).

139 Wolffe, A. & Pruss, D. Targeting chromatin disruption: Transcription regulators that acetylate histones. *Cell* **84**, 817-819 (1996).

140 Azuara, V. *et al.* Chromatin signatures of pluripotent cell lines. *Nature cell biology* **8**, 532-538, doi:10.1038/ncb1403 (2006).

141 Mikkelsen, T. *et al.* Dissecting direct reprogramming through integrative genomic analysis. *Nature* **454**, 49-55, doi:10.1038/nature07056 (2008).

142 Mikkelsen, T. *et al.* Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature* **448**, 553-560, doi:10.1038/nature06008 (2007).

143 Meissner, A. *et al.* Genome-scale DNA methylation maps of pluripotent and differentiated cells. *Nature* **454**, 766-770, doi:10.1038/nature07107 (2008).

144 Yildirim, O. *et al.* Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. *Cell* **147**, 1498-1510, doi:10.1016/j.cell.2011.11.054 (2011).

145 Sansam, C. G. & Roberts, C. W. Epigenetics and cancer: altered chromatin remodeling via Snf5 loss leads to aberrant cell cycle regulation. *Cell Cycle* **5**, 621-624 (2006).

146 Reisman, D., Glaros, S. & Thompson, E. A. The SWI/SNF complex and cancer. *Oncogene* **28**, 1653-1668 (2009).

147 Wang, J. *et al.* A protein interaction network for pluripotency of embryonic stem cells. *Nature* **444**, 364-368 (2006).

# Chapter 2 Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming

Chapter 2 was revised from the following publication to reflect the contributions of April Cook: "Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming" (2014) Nature Communications, in press. It can be examined as a final submission for publication in Appendix 1. My contribution to the work was generation of human data, and all author contributions can be seen in the Contributions section at the end of the chapter.

## Authors and affiliations

Jason A. West[1,5,7], April Cook[1,7], Burak H. Alver[2], Matthias Stadtfeld[3], Aimee Deaton[1], Konrad Hochedlinger[4], Peter J. Park[2,6], Michael Y. Tolstorukov[1,6,7], Robert E. Kingston[1,6]

[1] Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts, USA and The Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA.

[2] Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA.

[3] The Helen L. and Martin S. Kimmel Center for Biology and Medicine, The Skirball Institute of Biomolecular Medicine, Department of Cell Biology, New York University School of Medicine, New York, New York, USA.

[4] Howard Hughes Medical Institute and the Center for Regenerative Medicine, Massachusetts General Hospital, Boston, Massachusetts, USA and The Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts, USA.

[5] Present address: Therapeutic Innovation Unit, Amgen, Inc., Cambridge, Massachusetts, USA.

[6] Corresponding Authors: kingston@molbio.mgh.harvard.edu, tolstorukov@molbio.mgh.harvard.edu, peter_park@hms.harvard.edu

[7] These authors contributed equally to this work.

## Abstract

Chromatin structure is a fundamental determinant of DNA accessibility. Here, we examine nucleosome occupancy in mouse and human embryonic stem cells (ESCs), induced-pluripotent stem cells (iPSCs), and differentiated cell types using MNase-seq. To address variability inherent in this technique, we developed a bioinformatic approach that enabled the identification of regions of difference (RoD) in nucleosome occupancy between pluripotent and somatic cells. Most regions remain unchanged and, surprisingly, a majority of RoDs are the size of a single nucleosome. They are enriched at genes and regulatory elements, including enhancers associated with pluripotency and differentiation. RoDs correlate significantly with binding sites for regulators of development and pluripotency. We see extensive alterations of nucleosome landscapes in ESC enhancers, and observe lower nucleosome occupancy in these regulatory regions in pluripotent cells when compared to somatic cells. The majority of these changes in nucleosome signatures are reset during reprogramming. We conclude that changes in nucleosome occupancy are a hallmark of pluripotency and likely identify key regulatory regions that play a role in determining cell identity.

## INTRODUCTION

Embryonic stem cells (ESCs) and induced-pluripotent stem cells (iPSCs) self-renew and differentiate into an array of cell types *in vitro* and *in vivo*. A complex network of genetic and epigenetic pathways regulates the self-renewal and differentiation of these pluripotent cells, and the structure and covalent modifications of chromatin play a prominent role in this process. Prior work has established multiple unique properties of pluripotent chromatin and its regulation, including macrostructural descriptions of ESC chromatin as relatively “open”

compared to lineage-committed cells[1-6]. The pluripotency factors *Oct4*, *Sox2* and *Nanog* transcriptionally regulate and interact with certain chromatin-remodeling and histone-modifying complexes[7]. Reciprocally, multiple chromatin regulators have been implicated in the maintenance of pluripotency and during cellular differentiation and development, including a chromatin-remodeling complex unique to ESCs[1-3,8-10].

The physical packaging of DNA into nucleosomes, including the location of those nucleosomes on the genome, is known to be a central determinant of DNA accessibility in both *cis* and *trans*. Nucleosomes consist of approximately 150bp of DNA wrapped around a core histone octamer[11,12]. Nucleosome positioning is dynamic and can determine the ability of regulatory factors to bind, which impacts processes ranging from gene regulation to DNA replication, recombination, and repair[13,14]. Thus, characterizing changes in nucleosome occupancy is expected to reveal important regulatory features in pluripotent cell biology, differentiation, and reprogramming. These changes might be spread throughout the genome or localized to specific regions, and they might or might not reset completely upon reprogramming. Information on nucleosome location can be integrated with previous studies on covalent changes to chromatin (e.g., DNA and histone methylation, histone acetylation) to provide a more complete understanding of how chromatin dynamics contribute to pluripotency.

Development of techniques for mapping nucleosome positioning on the genome scale has illuminated the role of primary chromatin structure in the mammalian cell [15-22]. However, comparing the nucleosomal profiles between different cell types still presents profound challenges. Observed nucleosome occupancy is sensitive to even slight variations in

experimental conditions, such as the degree of chromatin fragmentation or the salt concentration used for chromatin isolation[23,24]. This variability is hard to control and, as a result, dynamic changes in nucleosome occupancy and positions associated with biological processes in mammalian cells have been difficult to quantify. In particular, it is not clear if large scale or local nucleosome occupancy changes are prevalent in these processes and how these changes contribute to alterations in gene expression.

Here, we investigate nucleosome occupancy within mammalian pluripotent and somatic cell populations and identify regions of differences between ESCs, iPSCs, and somatic cells in both human and mouse. This analysis is made possible by a novel data processing method developed for pair-wise comparisons of nucleosome occupancy measured in different conditions and cell types. We report that the observed differences are mostly the size of single nucleosomes, are enriched for motifs of transcription factors that drive pluripotency and somatic cell reprogramming, and reside within key regulatory regions of the genome, specifically at transcriptional start sites (TSSs) and enhancers of genes linked to pluripotency and differentiation. These findings reveal that localized changes in nucleosome occupancy at key regulatory regions, rather than large-scale rearrangements, may be sufficient to impact cell identity.

## Results

### Determining genome-wide nucleosome occupancy maps of human pluripotent and somatic cells

We profiled primary chromatin structure in three human cell types: H1OCT4GFP ESCs, iPSCs derived from fibroblasts that were differentiated from those H1OCT4GFP ESCs, and fibroblasts that were differentiated from the H1OCT4GFP ESCs. The three cell lines are isogenic, controlling for any effects of underlying sequence differences on nucleosome occupancy. The ESC and iPSC were described previously[25,26]. For each cell type, we created a nucleosome occupancy profile for the physical location of nucleosomes on their genomic DNA. To create these nucleosome occupancy profiles, we measured DNA protection patterns after chromatin digestion by micrococcal nuclease (MNase), building upon strategies previously developed by our group and others[15,17,20,27-30]. MNase selectively cleaves chromatin in linker DNA between nucleosomes, and sequenced mononucleosome-sized DNA fragments predict nucleosome occupancy in a given cell preparation. We generated over 100 million mapped paired-end reads for each cell type resulting in a detailed description of the nucleosome occupancy for the various pluripotent and somatic cell lines. The average fragment length from each library was near the predicted mononucleosome DNA fragment length (approximately 150 bp); though it should be noted that our samples are cross-linked and a fraction of DNA fragments could have been protected by DNA-binding factors rather than nucleosomes[31]. Conversely, due to the preferential elimination of longer fragments during library preparation and sequencing, our data set may be depleted of the nucleosomes bound by larger complexes such as PolII[32] or organized into heterochromatin. With these limitations in mind we use the term nucleosome

occupancy to characterize the number of digestion fragments at a given genomic position. Libraries showed high complexity with low percentages of repeats.

For comparison of results we also profiled primary chromatin structure in five murine cell types: mouse embryonic stem cells (mESCs), induced-pluripotent stem cells derived from tail-tip-fibroblasts (miPSC-TTFs) and liver (miPSC-Liver), somatic TTFs, and somatic liver. All cells originated from the same isogenic mouse line and have been previously characterized and described[33]. Importantly, the same trends were observed in the data derived from human and mouse samples. For more details, see our manuscript, Appendix 1.

We first assessed the average nucleosome occupancy patterns at the transcription start sites (TSSs) for each cell type. As demonstrated previously,[16,17,19,27] a nucleosome-depleted region (NDR) flanked by well-positioned +1 and -1 nucleosomes (relative to the TSS) is a characteristic feature of the occupancy profiles averaged across all genes. We observed a NDR flanked by well-positioned nucleosomes at the TSSs across all samples (Figure 2.1A,C). Despite this consistent pattern, we observed high variability in average nucleosome density, even for biological replicates from the same cell type and for ESCs and iPSCs (Figure 2.1A,C). Such variability is not specific to our experimental protocol since previous studies in mammalian genomes reported substantially different nucleosomal patterns at TSSs ranging from an accumulation in tag counts greater than the surrounding regions to an apparent depletion in occupancy[16-19,22,34]. This variability likely originates from technical rather than biological reasons, such as differences in experimental conditions and specifically the extent of MNase digestion, which is difficult to control between independent chromatin preparations and hinders direct comparisons of the nucleosome occupancy between cell types.

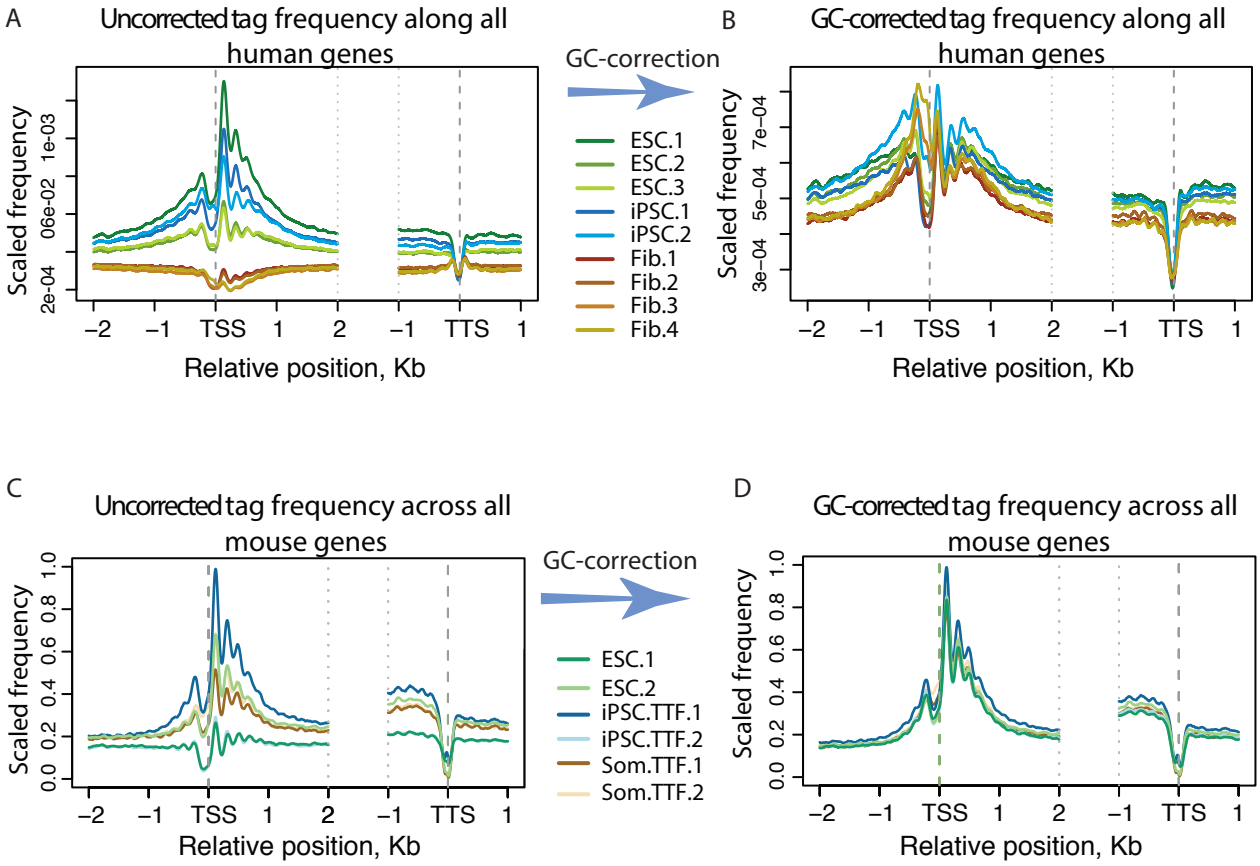



**Figure 2.1 Comparison of nucleosome occupancy in human pluripotent and differentiated cells**. **A.** Nucleosome occupancy around transcription start and end sites computed for human ESCs, iPSCs and differentiated fibroblasts (fib). We note that after normalizing the occupancy for the total number of tags in each library the profiles remain different, even between replicates that belong to the same cell type. **B.** The same profiles after normalization of the GC-content distribution in each sample with the target mean GC content of 50% (see Methods for more detail). **C,D** same as in **A,B** but with mouse data.

Among the characteristics of the MNase-seq data that correlate with the extent of MNase digestion is the GC-content distribution of the sequenced fragments, which was highly variable across all samples, including biological replicates. It is known that GC content of a population of size-selected fragments of MNase-digested DNA can change, in fact increase, with increasing digestion[35]. This is likely due in part to the MNase enzyme bias towards cutting AT-rich sequences. For an example of how the level of MNase digestion impacts both average fragment size and the GC content of the MNase-liberated fragments see Figure 2.2. We expect GC-content distribution of the mononucleosomal fragments to be constant between replicates because of careful control of digestion conditions, DNA fragment selection, and library preparation; however we see some variability. To address this issue, we included a step in our methodology that used GC-content of DNA sequence as a metric for normalization (Figure 2.3). Previously, nucleotide composition or GC content normalization has been applied to the analysis of microarray and high-throughput sequencing data (ChIP-chip, ChIP-seq, DNA-seq, etc)[36-38]. Here, we applied a concept similar to that used for ChIP-seq data[37] to the data produced by MNase digestion profiling. We normalized GC-content in each sample along the entire genome to a target value of 50%, which roughly corresponds to the average GC-content in the TSS-proximal regions in the mammalian genome (Figure 2.3). The GC-content normalization markedly reduced variability across all profiles at TSSs in both human and murine data (Figure 2.1B,D). Since TSS profiles are the result of averaging large sets of genomic loci, they should be similar for samples demonstrating similar gene expression patterns such as replicates of the same cell type. To directly evaluate the extent of similarity we computed

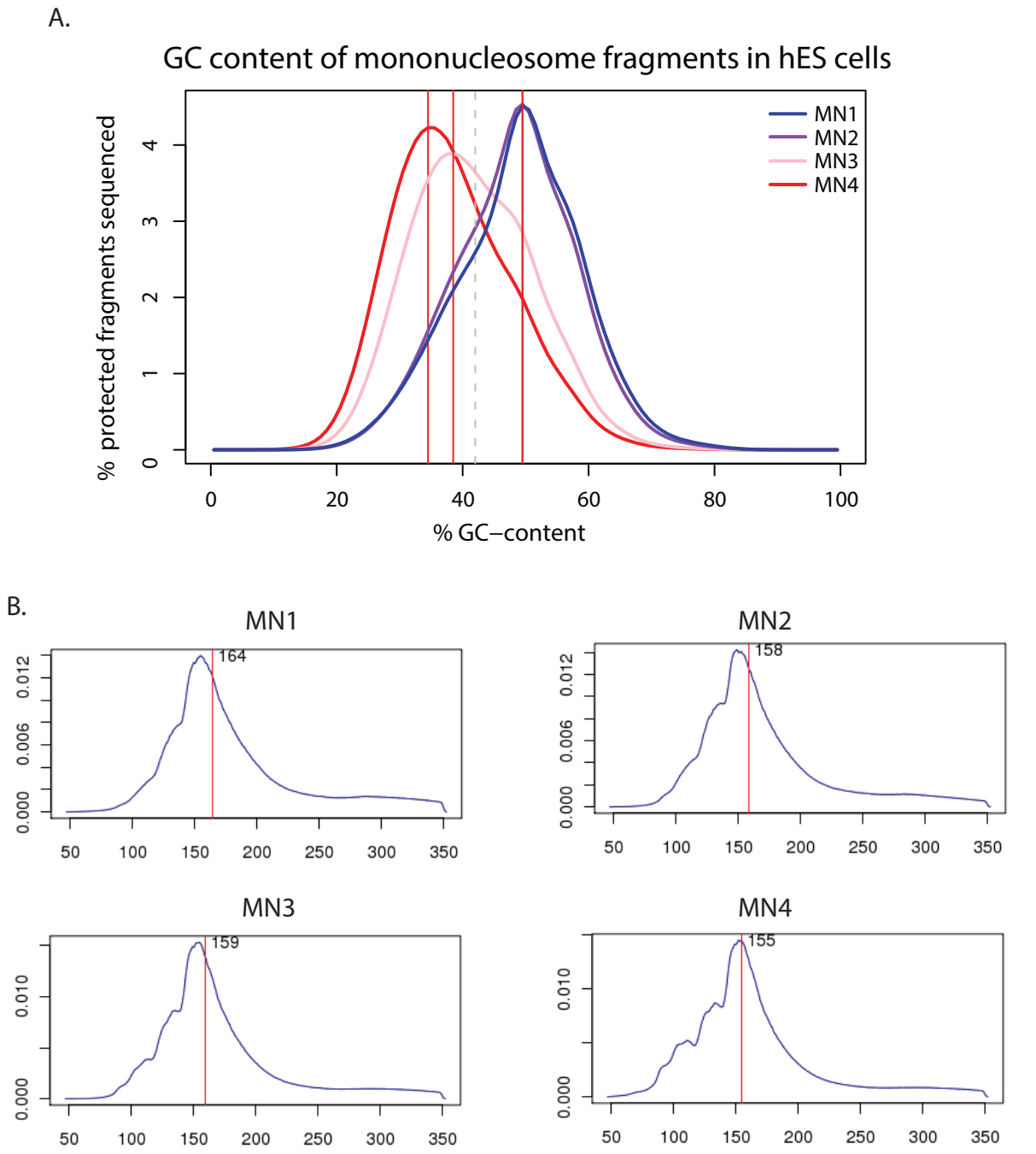



**Figure 2.2 GC content of sequenced fragments of DNA across a range of digestion correlates with fragment length and extent of digestion**. Each sample was digested with four concentrations of MNase, increasing from MN1 to MN4. The mononucleosome sized DNA from each condition was barcoded and sequenced. **A.** GC content of every fragment protected by nucleosomes is calculated, and percent of total fragments per GC content are plotted. **B.** Fragment distribution size for the nucleosome-protected fragments from each digestion condition.

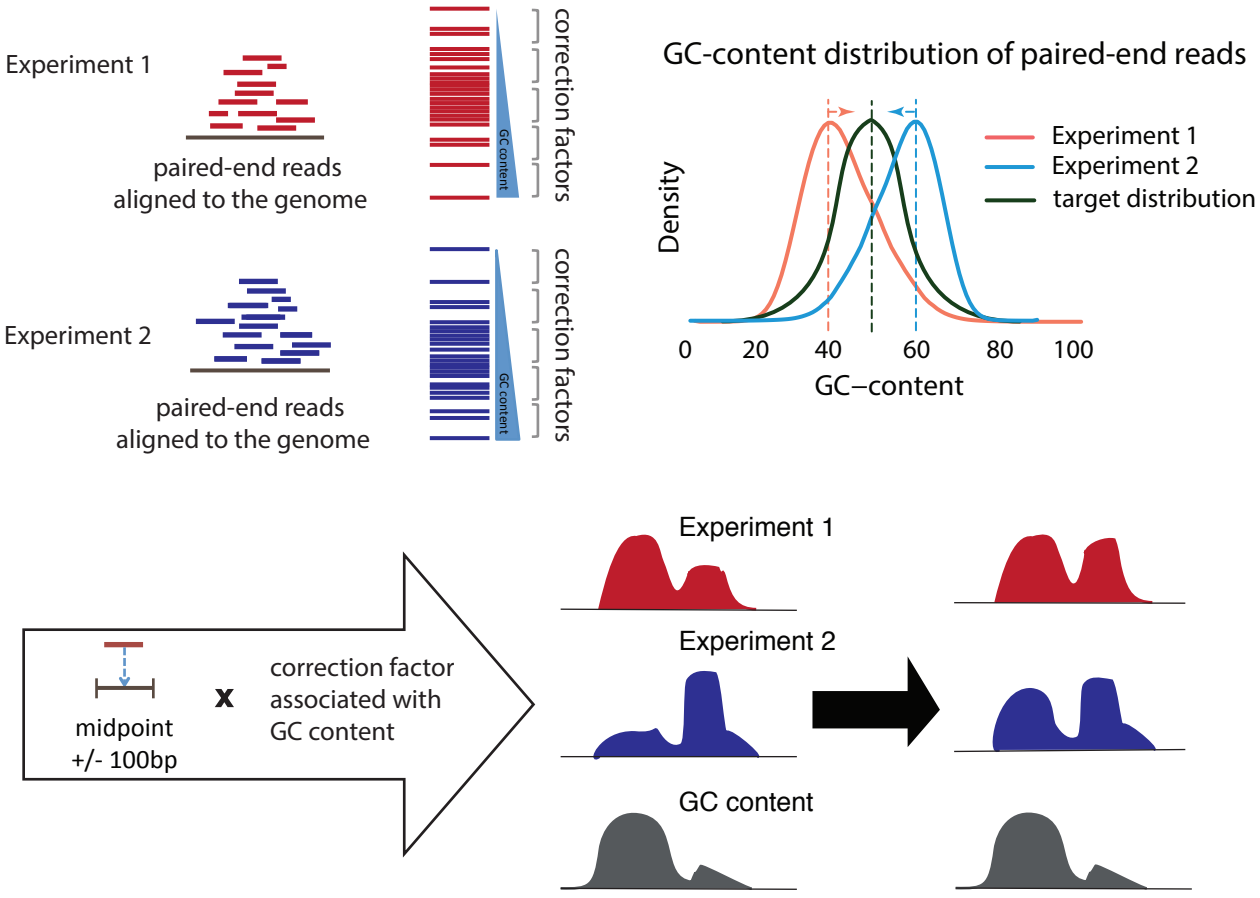



**Figure 2.3 Schematic illustration of the GC-correction procedure applied to MNase-Seq data in this study**. Two libraries generated in the independent experiments are often characterized by distinct distributions of GC-content of the fragments. This results in biased tag frequency profiles. To correct for this effect we compute the correction factor for each GC-content value in such a way that the actual GC-content distributions 'mimic' target distribution. These correction factors are applied to tag frequencies at each genomic position based on the GC-content of the locus encompassing it and generate the corrected tag frequency profiles that are used in for further analysis.

correlation of nucleosome density at TSSs in mouse (measured as average normalized frequency of fragments

per kilobase of DNA) and observed increased correspondence between replicates of the same cell type upon GC-normalization (Figure 2.4).

We note that the magnitude of the nucleosomal signal at TSSs detected in a particular study depends on a number of factors, including nucleosome stability, accessibility and turnover rate. For example, using different salt fractions during chromatin isolation results in different TSS-proximal MNase-seq profiles[24]. Similarly, different levels of MNase digestion can produce TSS-proximal profiles with different shapes, each reflecting nucleosome occupancy as determined by a specific set of sample preparation conditions. Therefore, to further validate our results, we assessed another target GC (48%, which is equal to the average GC-content of our mouse samples) confirming that our conclusions are not limited to a specific target GC-content used for normalization. Thus, we conclude that the GC-normalization effectively reduces variability present in MNase-seq data sets and enables comparisons of nucleosome occupancy across different cell types. Equipped with this methodology, we proceeded to identifying defined changes in nucleosome occupancy in pluripotent and somatic cells.

## Chromatin structure changes at regulatory loci

We began by investigating differences in nucleosome organization at gene promoters and enhancers where we hypothesize it to play a regulatory role, and then extended the analysis to the whole genome. Enhancers are a class of regulatory regions key for the pluripotent state. Here we used recently published sets of enhancers showing strong association with the pluripotency and reprogramming factors Oct4, Nanog, and Sox2, including

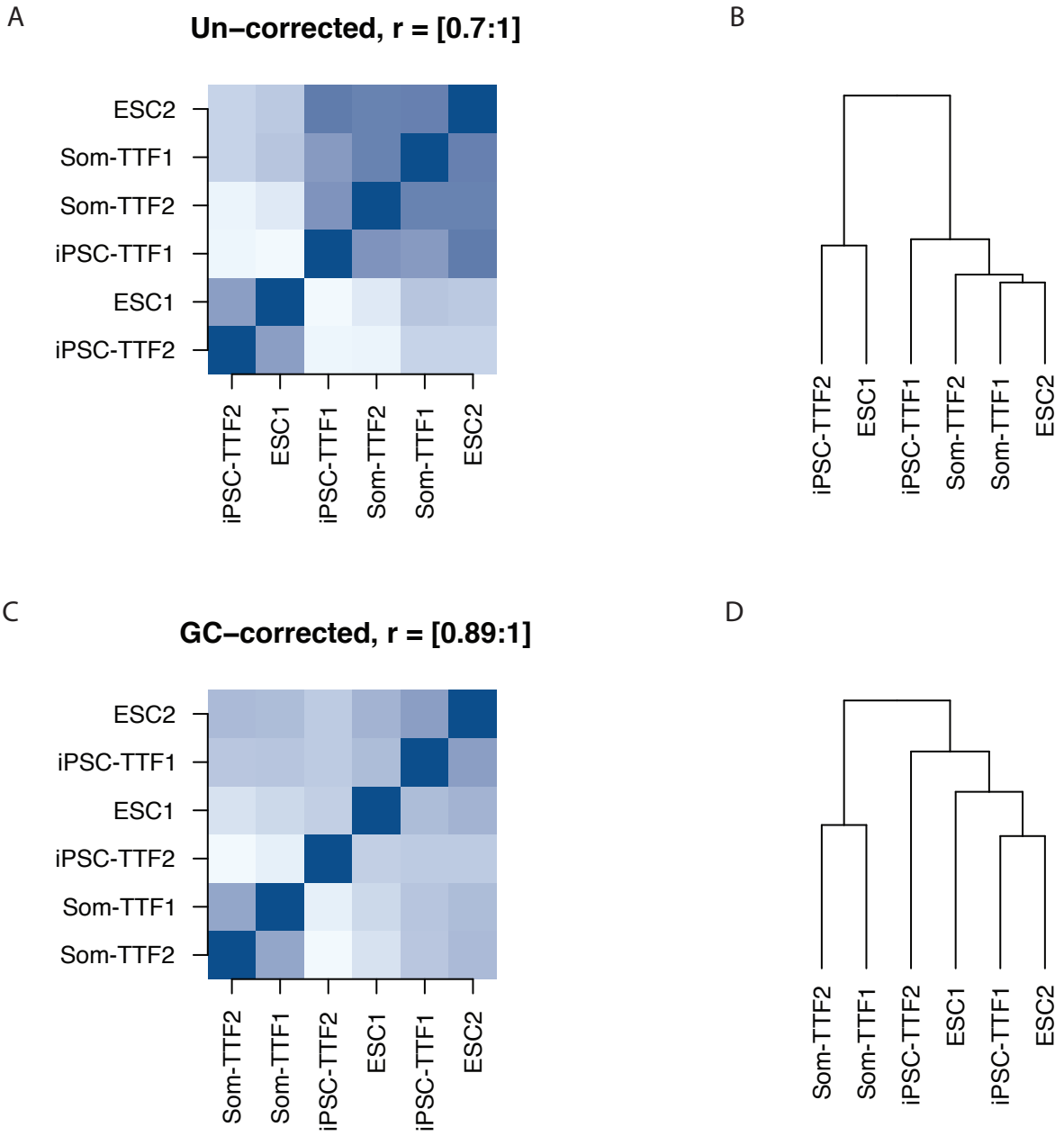



**Figure 2.4 Effect of GC-correction on replicate similarity.** The results are based on correlation between the normalized tag counts in TSS-proximal regions (+/-2kb) in each individual replicate of mES, iPS-TTF, and somatic TTF cells. The results are shown separately for GC-uncorrected **A,B** and GC-corrected **C,D** counts. Panels **A,C** represent heat maps (dark blue is high correlation coefficient, light blue is low) and panels **B,D** represent clustering dendrograms. We note that GC-correction results in placement of somatic replicates in the same cluster and in better correspondence between replicates of pluripotent cell types.

a subset of 'super-enhancers' that are unusually large and impart hyper-regulatory functions in ESCs[39,40]. The set comprises 7006 human ES cell enhancers, 684 of which are super-enhancers. Comparison of the nucleosome occupancy profiles around scaled ESC enhancers in somatic and pluripotent cells revealed that on average the occupancy was lower in pluripotent cells than in differentiated cells (Figure 2.5A), which is consistent with these regions being more accessible to regulatory proteins in pluripotent cells. The same trend was observed in mouse MNase-seq data for mESCs, miPSCs, and tail-tip fibroblasts (Figure 2.5B).

For a more detailed analysis, we used mouse MNase-seq and enhancer data and divided enhancers, (n= 8794, 231 of these are 'super enhancers') into two groups, those having significantly lower and higher nucleosome density in ESCs when compared to differentiated fibroblasts (LND and HND groups; significance estimates were based on the variability of the nucleosome density in the available replicates, see Methods). In line with the results discussed above, the LND group comprised 353 enhancers (23 of which were super-enhancers), while the HND group comprised only 60 enhancers (one of which was a super-enhancer). When all TSS-proximal regions were similarly divided into LND and HND groups for comparison, the corresponding counts were 558 and 341, thus resulting in considerably less skewed group counts than were detected at enhancers.

The functional importance of nucleosome occupancy change at mouse enhancers was further substantiated by the gene ontology (GO) analysis, which revealed that the genes associated with LND enhancers were enriched in such terms as 'cell and tissue differentiation', 'embryo and epithelium development', and 'regulation of transcription from RNA polymerase II

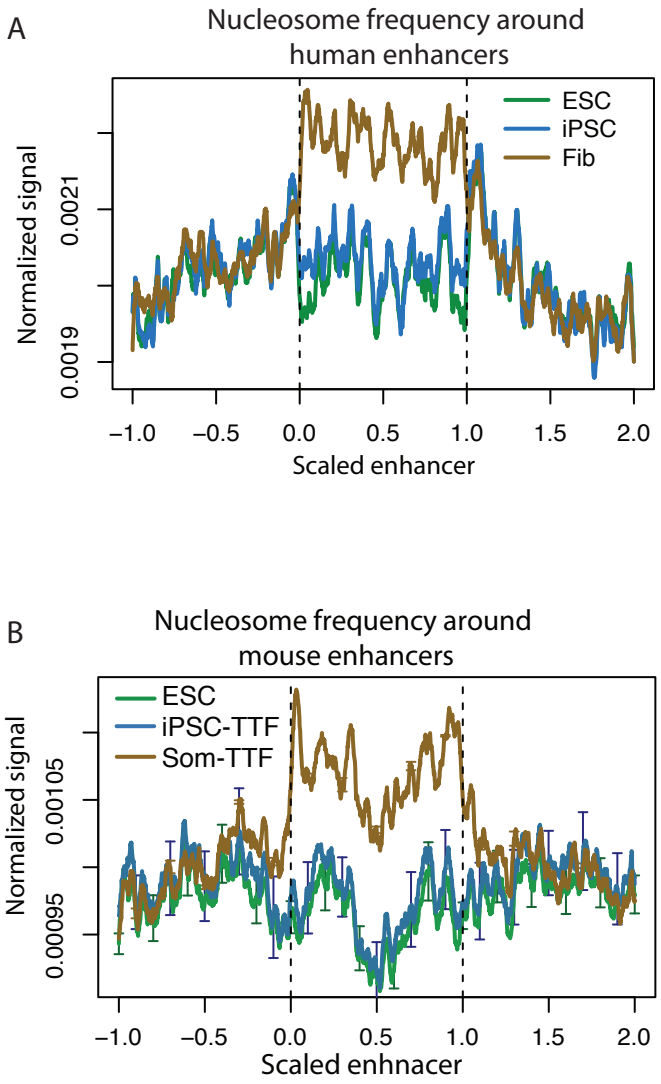



**Figure 2.5 Comparison of nucleosome occupancy at enhancers in pluripotent and differentiated cells. A, B**. Normalized nucleosome occupancy signal around scaled ESC enhancer regions computed for replicate sets in three cell types for human, **A.**, and mouse, **B.**

promoter'. The genes associated with HND enhancers did not exhibit significant enrichment in any GO term, except for 'unannotated'.

To further investigate how nucleosome occupancy at enhancers correlates with other features of chromatin organization, we used published data on a number of chromatin marks and other data types at these regions in mouse[39]. Enhancers with LND were more likely to be bound by transcription factors (Oct, Klf4, Sox2), exhibited active chromatin marks, and were associated with stronger DNase I signal when compared to enhancers from the HND group (Figure 2.6). This rearrangement of the nucleosome landscape at enhancers might be a key determinant in pluripotency and differentiation, with lower nucleosome occupancy correlating with stronger enhancer activity in pluripotent cells. We conclude that the rearrangement of nucleosome landscape at regulatory regions correlates with changes in other chromatin signatures in a cell type-specific manner, and that active enhancers show lower levels of nucleosome occupancy in pluripotent versus differentiated cells.

## Genome-wide comparison of pluripotent and somatic cells reveals punctate regions of difference in nucleosome occupancy at key regulatory regions

To expand the analysis of changes in nucleosome occupancy, we sought to identify all regions of difference (RoD) in the nucleosome occupancy profiles of ESCs, iPSCs, and differentiated cells on a genome scale, regardless of their location relative to annotated DNA elements. Nucleosome organization is likely to undergo re-arrangement as cells change fate, and visual inspection of the nucleosome occupancy profiles revealed such changes (Figure 2.7A-C). However, little is known about the nature of nucleosome occupancy changes on the genomic scale, including their significance, prevalence, size, and distribution, in part due to the

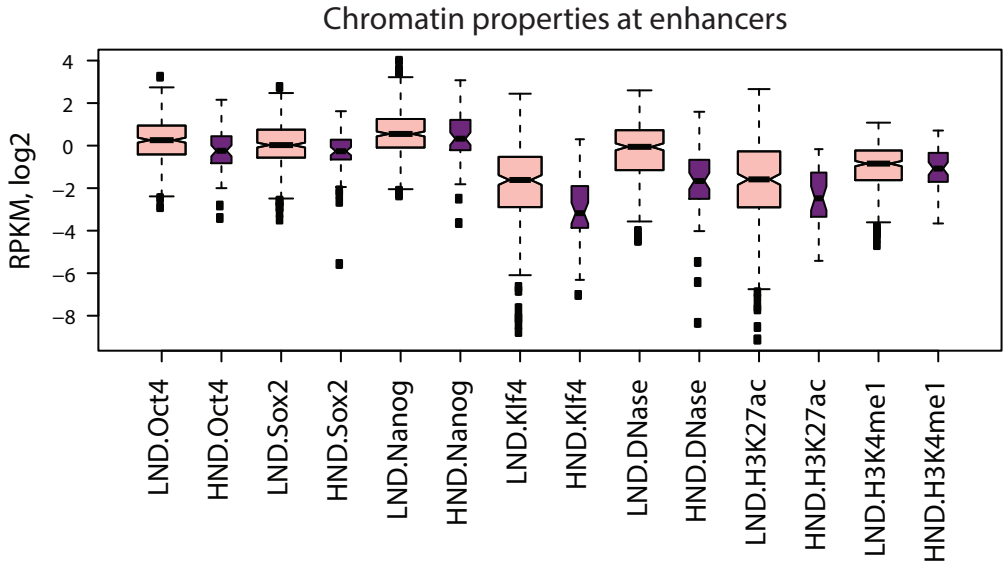



**Figure 2.6 Comparison of classes of nucleosome occupancy in mouse pluripotent and differentiated cells..** Chromatin property comparison (measured in mESCs in Whyte, 2013) for the LND and HND enhancers. Notches provide 95% confidence intervals.

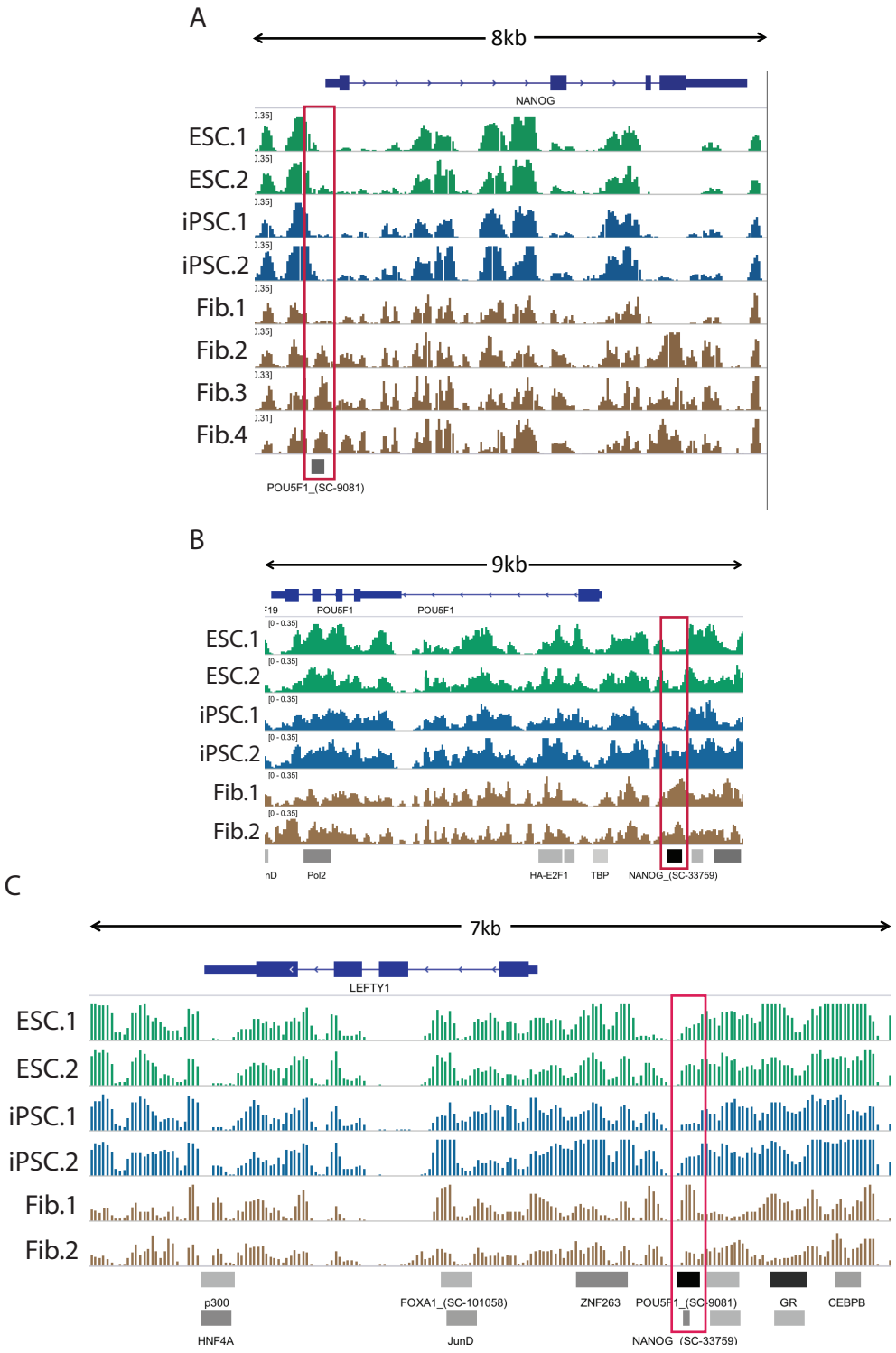



**Figure 2.7 Screenshots of regions in the human genome where occupancy differs in pluripotent and differentiated cell types. A.** An OCT4 binding site at the NANOG TSS shows low occupancy in pluripotent cells where this gene is active, and occupancy in differentiated fibroblasts where it is not active. **B.** A NANOG binding site is open in pluripotent and occupied in fibroblasts at the OCT4 gene. **C.** A OCT4/Nanog binding site upstream of the LEFTY1 gene where occupancy is increased in differentiated cells with respect to pluripotent cells.

challenges inherent in mapping these differences in mammalian cells. We applied a novel approach comparing the nucleosome density in 150-bp bins to scan the genome and generate p-value profiles describing the significance in nucleosome occupancy differences between pluripotent and somatic cell lines (Figure 2.8A). We note that since this algorithm is not focused on stable nucleosome positions it is suitable for detection of RoDs of any size, and using a false discovery rate (FDR) threshold we were able to identify significant RoDs in pairwise cell-type comparisons (see Methods for details).

GC normalization, one of the features that distinguish our approach from earlier algorithms[41], facilitated the identification of RoDs by reducing variability between replicates. This allowed identification of approximately 45% more RoDs in the comparison of mESCs and somatic TTFs than in non-normalized data. To evaluate the extent to which somatic cell reprogramming resets the chromatin structure in iPSCs, we compared the numbers of RoDs identified between pluripotent and differentiated cell types with those detected between ESCs and iPSCs. As the number of detected RoDs is a function of the selected significance threshold, we analyzed RoD counts for a series of thresholds in GC-normalized data. We consistently identified more RoDs in pluripotent versus differentiated cell comparisons than comparisons of two independent pluripotent cell lines in both mouse and human data (Figure 2.7B,C). For instance, at FDR=0.1, we identified more than 100,000 RoDs when hESCs were compared to fibroblasts, and over 200,000 RoDs when hiPSCs were compared to fibroblasts. For the hESCs and hiPSC comparison, 12,000 RoDs were identified, which is more than eight fold lower than the number of RoDs identified in any pluripotent versus somatic cell comparison. ESCs and iPSC

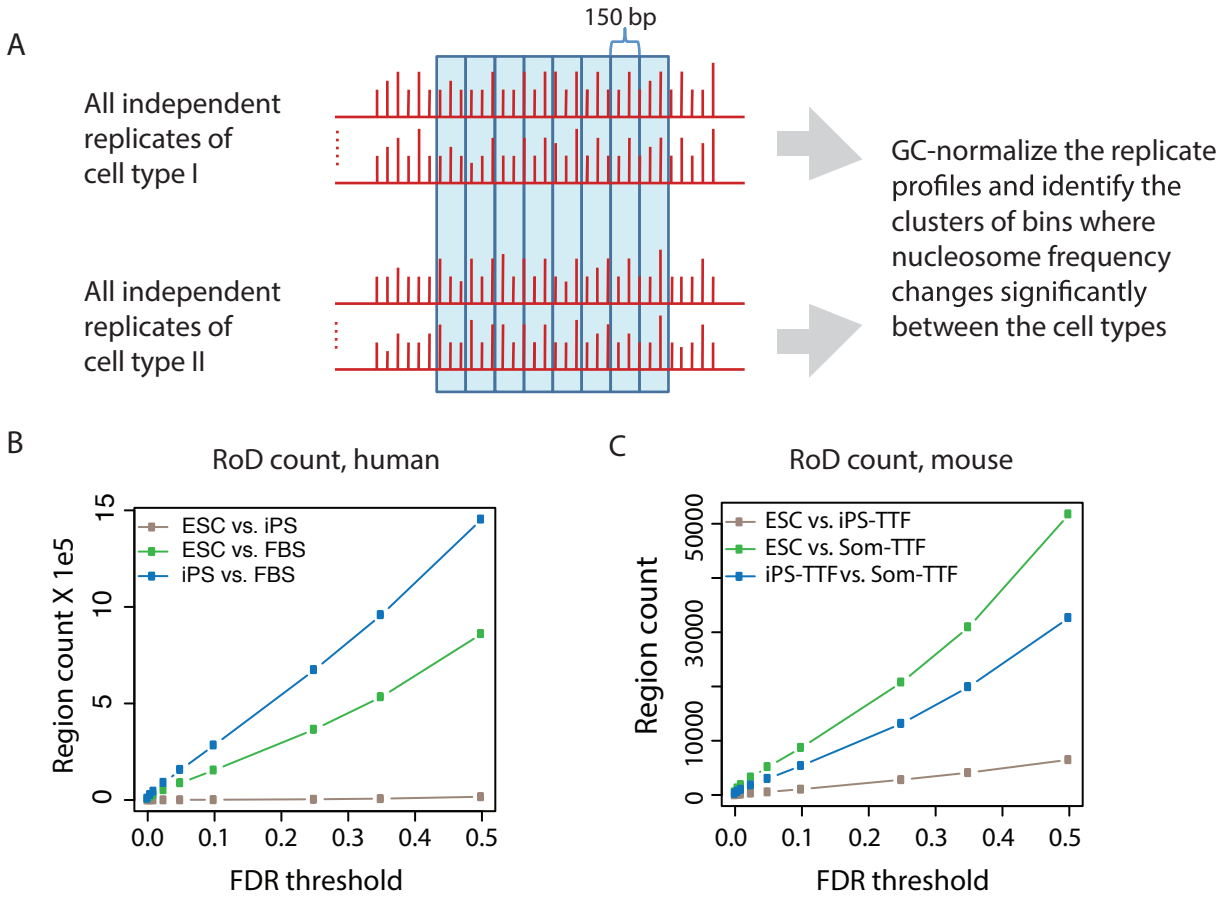



**Figure 2.8 Identification of regions of difference (RoDs) in nucleosome occupancy profiles between pluripotent and differentiated cell types. A.** Schematic illustration of the method used for RoD identification. Sequenced tag frequencies in all replicates of the compared cell types (red) were binned along the genomic coordinate (blue) and the clusters of bins where tag frequencies were significantly different were retained for further analysis (see Methods for detail). . B, Human, C, mouse counts of the RoDs identified with different false discovery rate thresholds (FDR=0.1 was selected for representative RoD sets for downstream analyses)

are known to have a very similar transcriptional profile[42], which is consistent with the lower number of RoDs detected when comparing these cell types.

While transcriptional similarities are well studied between ESC and iPSC lines, one hypothesis that remained untested was that iPSCs could more closely resemble their cell of origin rather than ESCs with regard to nucleosome positioning. However, based on previous work, ESCs and iPSCs are functionally equivalent and very similar at the molecular level (reviewed in [43]). Thus one would anticipate a high degree of similarity between iPSCs and ESCs in nucleosomal occupancy profiles. Indeed, the differences in nucleosome organization observed in the comparisons of somatic cells to ESCs correlate with the differences detected in comparisons with iPSC (Figure 2.9A,B). For instance, all the regions determined for a selected FDR threshold in hESCs exhibit the same directional change in the hiPSC comparison, and vice versa (green and blue dots in Figure 2.9A). These observations were further confirmed in mESC, miPSC, and TTF comparisons (Figure 2.9B).

We also examined two basic characteristics of RoDs: their size distributions and the direction of nucleosome occupancy change. Surprisingly, the vast majority of RoDs were 150 bp in size (more than 95% in both the human and mouse data), suggesting tight control of chromatin structure at the level of single nucleosomes. A small number of RoDs were several kilobases in length, but no regions larger than 10kb were observed (Figure 2.10A). When directionality of the occupancy change was considered, the majority of the RoDs identified when comparing pluripotent cells to differentiated cells showed an increase in nucleosome signal in differentiated cells (Figure 2.10B,C). This supports the hypothesis that pluripotent cells

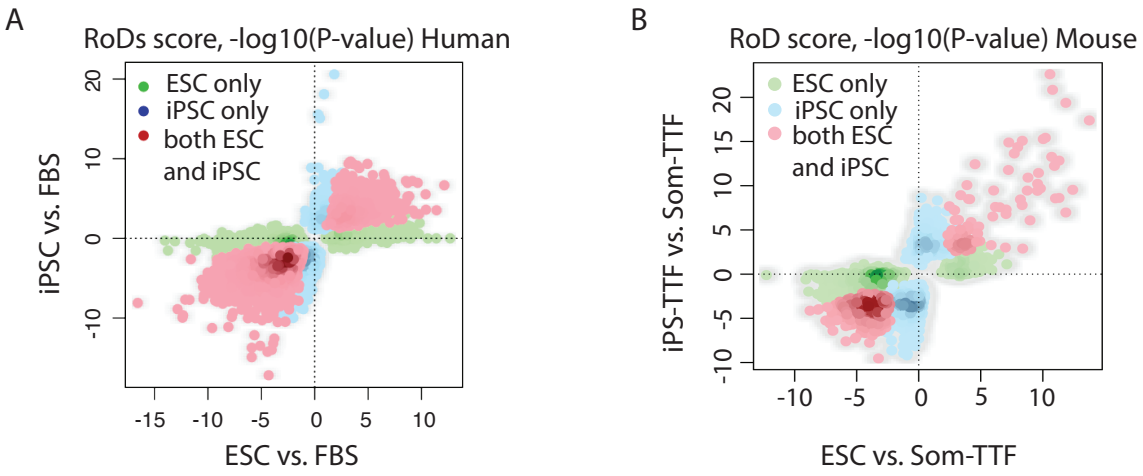



**Figure 2.9 Characterization of regions of difference. A.** Correlation between difference scores of the RoDs identified in comparisons of hESCs vs. fibroblasts and hiPS vs. fibroblasts, or in **B.** mESCs vs. somatic TTFs and miPS-TTFs vs. somatic TTFs. Only the bins that meet the FDR threshold of at least 0.1 in one comparison were taken for this analysis. Red dots represent bins that meet the selected FDR threshold in both comparisons; blue and green dots represent bins that meet the FDR threshold only in the 'iPSC vs. fibroblast' set or only 'ESC vs. fibroblast' set, respectively. We note that the sign of the score is maintained across the sets (i.e. bins that have positive (negative) scores in one pairwise cell-type comparison have the same score signs in the another pairwise cell-type comparison), which indicates of good correspondence between the loci of nucleosome occupancy variation in ESCs and iPSCs in both mouse and human data sets.

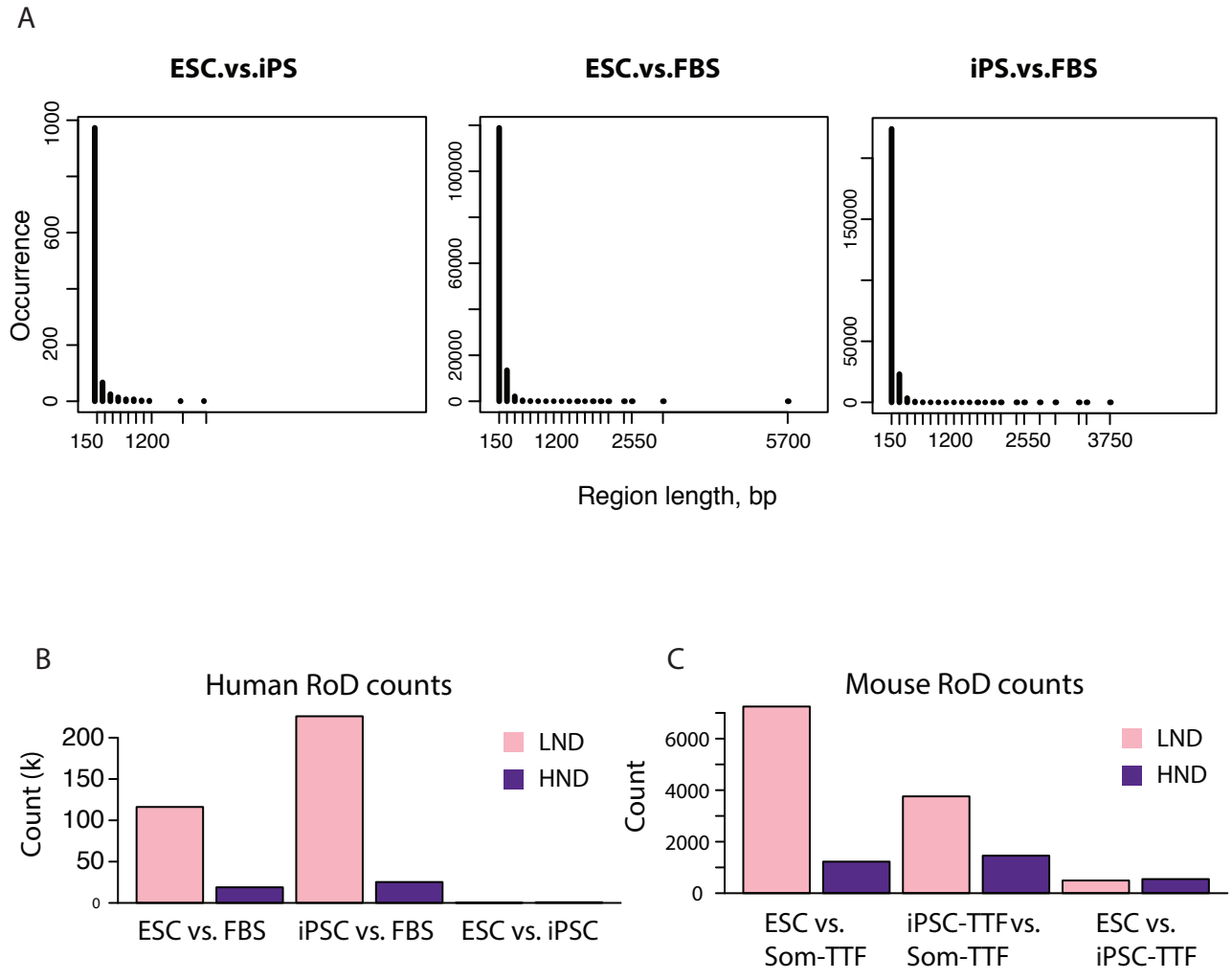



**Figure 2.10 Statistics on the regions of difference (RoDs) identified in pairwise comparisons of human and mouse cell types. A.** The distributions of lengths of the RoDs identified in the pair-wise comparisons of the human ESC versus iPSC, ESC versus fibroblast and iPSC versus fibroblasts. **B, C.** Occurrences of Human, **B.**, and mouse, **C.**, cell types. Comparison of the counts of RoDs with lower (pink) and higher (purple) levels of nucleosome occupancy in the pluripotent cell types relative to somatic TTFs (first two bar groups) and in ESCs relative to iPSCs (last bar group).

have relatively open chromatin, as one criterion for open chromatin would be lower nucleosome occupancy.

Thus our analysis revealed sets of the mostly punctate differences in nucleosome occupancy between pluripotent and differentiated cells. These loci are predominantly associated with lower nucleosome occupancy in the pluripotent cells. Overall, ESCs and iPSCs display a high degree of similarity in nucleosomal signal, providing evidence that somatic cell reprogramming into iPSCs resets nucleosome positioning to a pluripotent state[44]. We next sought to more fully characterize RoD locations, as these regions are likely regulatory sites involved in pluripotency and reprogramming.

## Regions of difference are enriched at regulatory regions active in mESCs

Our analysis showed that approximately 40% (42% in human) of the RoDs are at gene regions annotated in the mouse genome (Figure 2.11A-C), which is significantly more than expected for a randomized distribution of the RoDs in mappable regions of the genome ($P = 10^{-12}$, see Methods for details on significance estimation). Around genes, TSS proximal regions are specifically enriched in mouse RoDs (Figure 2.12, blue lines), including the promoters of genes associated with pluripotency and transcription activation. Indeed, in mouse pluripotent versus somatic cell comparisons, between 7 to 16% of RoDs occur at TSSs, and these are enriched 2.4 to 5 fold over the genome average (Figure 2.11 B,C). In addition to genes and their promoters, pluripotency-associated enhancers exhibited significant enrichment for mouse RoDs (Figure 2.12, orange lines). This was also the case for human RoDs (Figure 2.13 A). For a screenshot of the human data illustrating both the overall correlation between RoDs and genes as well as a specific enhancer containing region see Figure 2.13 B. Additionally, to our surprise, enhancers

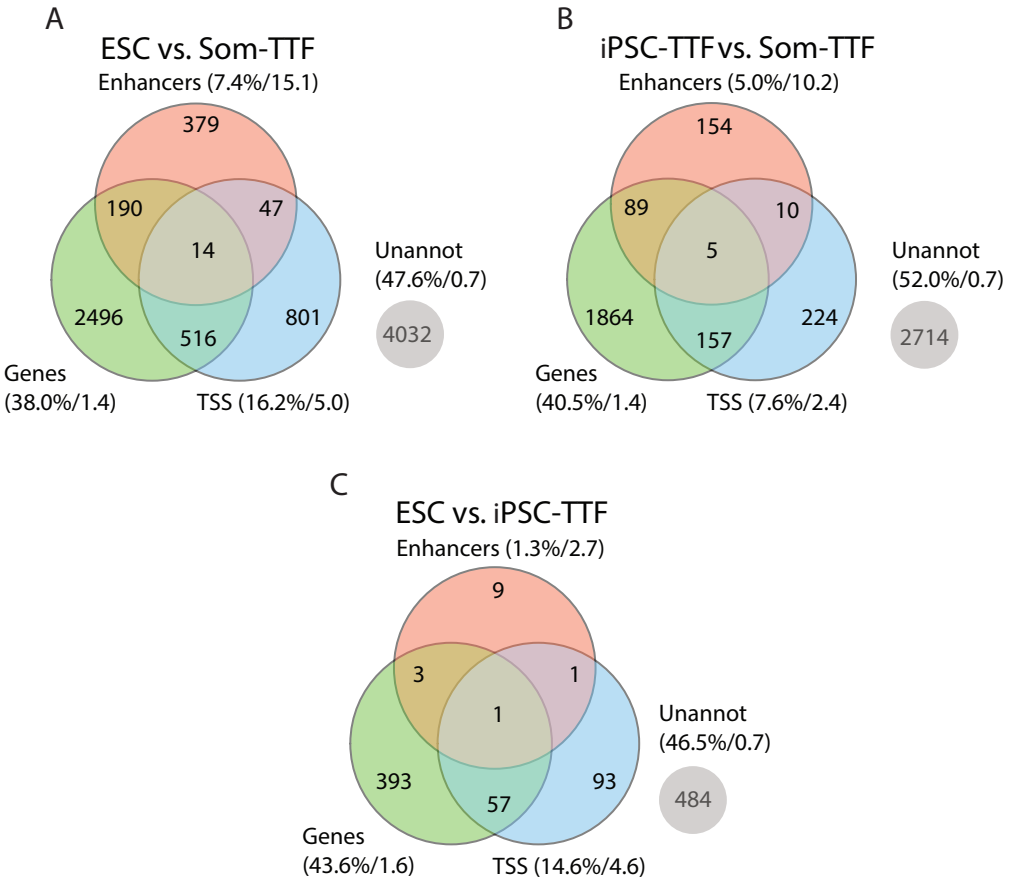



**Figure 2.11 Occurrences of the regions of difference (RoDs) identified in pairwise comparisons of mouse cell types**. **A-C.** Occurrences of the identified RoDs in the different regions of mouse genome for pair-wise comparisons of ESCs vs. somatic TTFs **A.** miPS-TTFs vs. somatic TTFs **B.** and ESCs versus miPS-TTFs **C.** Genes are defined according to USCS annotation for mm9 genome, TSS proximal regions comprise +/-2 Kb around gene starts, and ESC enhancer coordinates were taken from a recent publication. The numbers inside the circles represent counts of RoDs in corresponding regions. The numbers next to the region name represent the percentage of the RoD occurrences in this region to the total RoD count and the enrichment of this percentage over the expected value based on the region size in the genome. We note that the evaluated regions can overlap and therefore the sum of the percentages is not equal to 100%. This figure only includes RoDs meeting a false discovery rate threshold of 0.1.

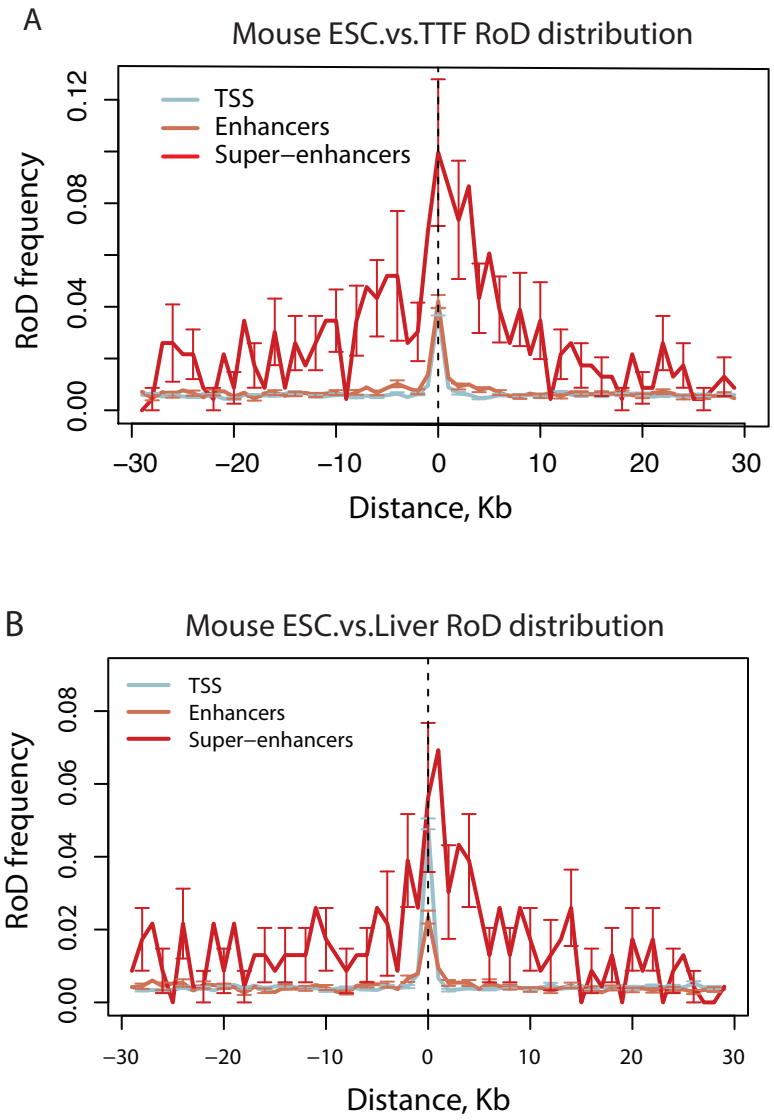



**Figure 2.12 The RoD frequencies in the regions encompassing transcription start sites and enhancers.** Enhancers identified in mESCs (Whyte 2013) **A. ESC** versus TTFs and **B.** ESCs versus Liver. The 95% confidence intervals are shown with the vertical arrows. The confidence intervals were estimated based on the variability of the frequency values in individual profiles used for averaging.

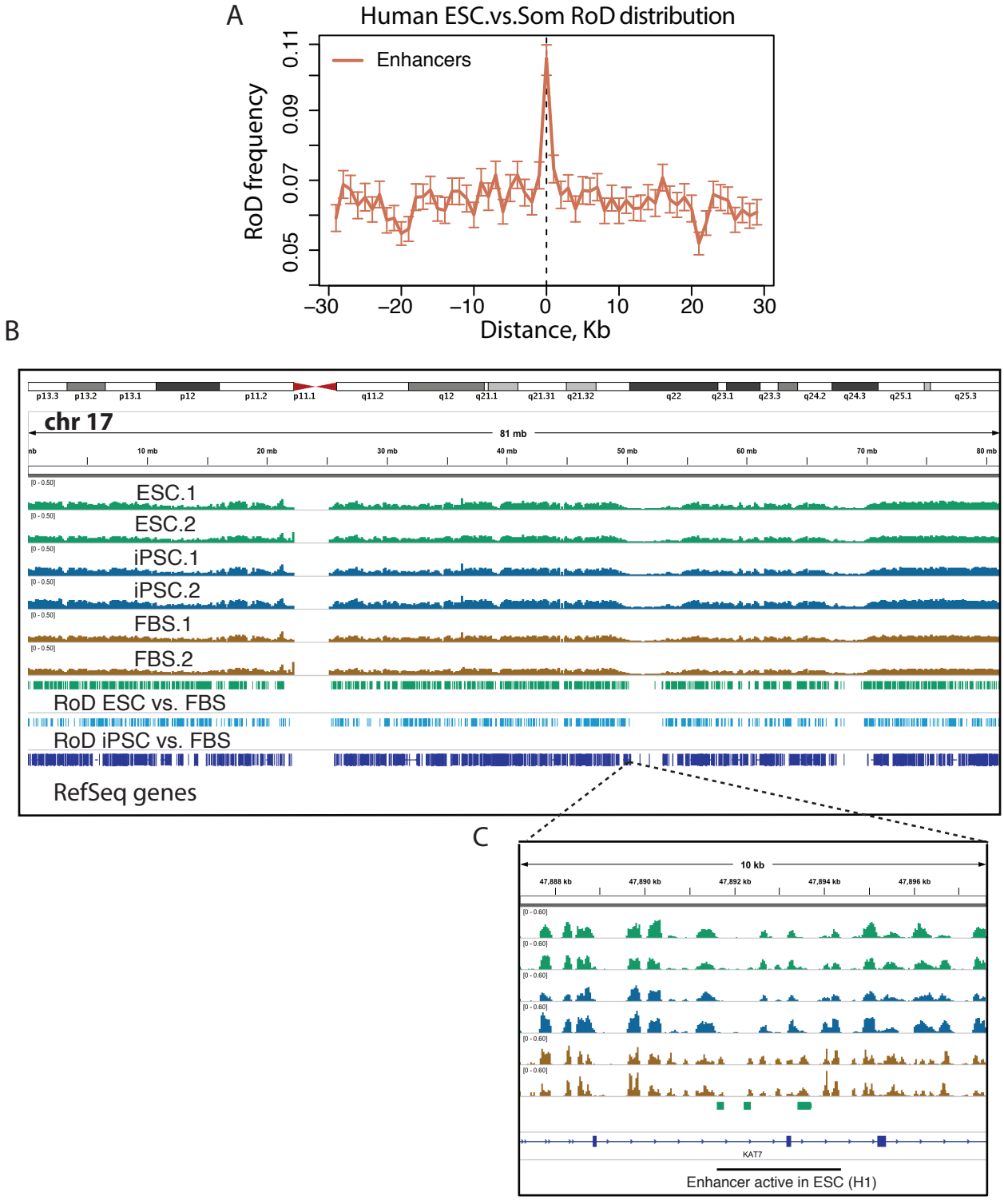



**Figure 2.13 Distribution of the regions of difference (RoDs) detected in nucleosome occupancy profiles relative to enhancers in the human genome. A.**RoD frequencies in regions encompassing enhancers identified in hESCs. **B.** Chromosome wide snapshot of nucleosome occupancy and RoD occurrence. **C.** Nucleosome occupancy at an enhancer identified in Hnisz et al., 2013

demonstrated differences with the same or greater magnitude as TSSs (which are known to show large differences) looked at in mouse. In the mouse pluripotent versus somatic cell comparisons, between 5 to 7.4% of RoDs occurred at mESC-defined enhancers, which corresponds to a 10 to 15 fold enrichment over what is expected at these enhancers (Figure 2.11 B,C). 'Super-enhancers' – large enhancer regions associated with a high density of regulatory protein binding[39] –showed an even stronger enrichment in mouse RoDs (Figure 2.12, red lines). As an additional validation of this result we identified RoDs between mESCs and another somatic cell type, mouse liver. This set of RoDs was also skewed towards LND enhancers in ESCs and showed enrichment at TSS and ESC enhancers (Figure 2.12B), confirming that these effects are not specific for somatic cell type to which ESCs are compared.

To further quantify the overlap between RoDs and these regulatory regions we computed the percent of mouse enhancers and TSSs harboring RoDs. We note that actual values of such an overlap would depend on the depth of the sequencing reached in a particular study (and thus the ability of the analysis to identify all nucleosome RoDs and enhancers) and the significance threshold used to call RoDs at a given locus. Under the threshold used in this study we found that 7% of the 'regular' enhancers and 39% of super-enhancers bear at least one RoD, which is moderate in value but represents a significant overlap as compared to the expected value for randomized RoD distribution ($P = 10^{-11}$, see Methods). A similar fraction of TSS proximal regions (6%) harbor RoDs, which reinforces the importance of chromatin structure and its regulation at enhancers in pluripotent and somatic cells. While most enhancers harbor only one or no RoDs, some, specifically super-enhancers, are associated with multiple RoDs.

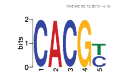
## Regions of difference are enriched for TF binding motifs associated with reprogramming

Given that RoDs are small in size (approximately 150 bp) and enriched at regulatory sites, one could hypothesize that they are associated with a regulatory protein binding events that displace a single nucleosome. For instance, regions associated with binding of TF involved in cell differentiation were reported to have lower nucleosome occupancy in the corresponding somatic cell type[21].

We focused on the RoDs that have lower nucleosome occupancy in pluripotent cells (LND RoDs) and analyzed them for the presence of sequence motifs to identify potential regulatory factors. We found that mouse LND RoDs identified in ESC versus somatic cell comparisons are enriched in motifs of transcription factors associated with reprogramming and pluripotency, including Klf4, c-Myc, Oct4, and Stat3 (Figure 2.14, mouse, and comparable analysis for human, Figure 2.15). As Oct4 and Sox2 act a heterodimer in pluripotent cells[45-47], we conclude that our analysis identifies the potential sites of functional binding for all four Yamanaka reprogramming factors in mouse. The Stat3 binding motif is also highly enriched in the RoD data, and Stat3 is required and sufficient for the self-renewal of mESCs[48]. Performing a *de novo* motif search with a random set of genomic sequences the same size as the mouse RoD set did not reveal motifs for the Yamanaka factors (with the selected significance threshold of E-value=$10^{-5}$). We note that many of the factors associated with the motifs identified in RoDs also bind enhancer regions in pluripotent cells and, furthermore, their binding is often used to define enhancer regions in pluripotent cells[39,47].

| Discovered motif | E-value | Known or similar motif |
|---|---|---|
|  | 8e-127 | Klf4, Klf7, Ascl2 |
|  | 1e-82 | Stat3, Esr1 |
|  | 6e-78 | Zfp281, Sp1, Zfp740 |
|  | 4e-54 | Rreb1 |
|  | 1e-54 | Ctcf, Gabpa, Ixr1 |
|  | 4e-46 | Ptx1, Bcd, Oc |
|  | 2e-24 | Sp1, Cha4, Zfp410 |
|  | 1e-24 | Eomes, Sna |
|  | 2e-20 | Unknown |
|  | 1e-12 | Myc, Max, Myc:Max |
|  | 8e-12 | Unknown |
|  | 3e-11 | Unknown |
|  | 5e-8 | Foxa1, Foxk1, Foxl1 |
|  | 2e-6 | Oct4 (Pou5f1), Pou2f2, Pou2f3 |
|  | 2e-6 | Elk4, Gabpa |

**Figure 2.14 Sequence motifs, mouse.** The complete list of sequence motifs found in de novo enrichment analysis of the RoDs associated with lower nucleosome density in mESCs compared to somatic TTF cells. Corresponding E-values and transcription factors associated with similar motifs are indicated.

| Discovered motif | E-value | Known or similar motif |
|---|---|---|
|  | 1.6e-135 | Zfp161/Max * |
|  | 7.5e-144 | Mtf1/Zfp105/Elf3 |
|  | 2.7e-120 | Zfp281 * |
|  | 4.9e-108 | Zbtb3 |
|  | 7.6e-085 | REST/TLX1::NFIC |
|  | 6.5e-013 | Pitx3/Hoxd1/Evx2 |
|  | 8.7e-065 | INSM1 |
|  | 1.0e-063 | Zfp691 |
|  | 2.6e-053 | MEF2A |
|  | 9.6e-039 | IRF1 |
|  | 9.6e-036 | ZEB1 |
|  | 2.0e-032 | Tcfcp2l1 |
|  | 5.7e-031 | Hic1 |
|  | 1.9e-028 | Klf4/Sox13/Ascl2 * |
|  | 2.2e-026 | FEV/Spi1/Stat3 * |
|  | 1.5e-018 | Hoxc6 |
|  | 3.7e-018 | HIF1A::ARNT/Mycn * |
|  | 6.2e-015 | Glis2 |
|  | 6.0e-014 | Sox7/Gm397 |

| Discovered motif | E-value | Known or similar motif |
|---|---|---|
|  | 4.6e-013 | RREB1/Egr1 * |
|  | 3.0e-012 | Irf4 |
|  | 7.7e-011 | Zfx/Egr1/TFAP2A |
|  | 1.4e-008 | Foxk1 * |
|  | 1.3e-007 | Fos/AP1/Jundm2 |
|  | 1.9e-007 | Hbp1 |
|  | 6.6e-007 | Zscan4 |
|  | 6.6e-007 | Zfp128/Six6 |
|  | 1.2e-005 | NFATC2 |
|  | 9.0e-004 | SPI1/Gabpa * |

similar/same in mouse *

**Figure 2.15 Sequence motifs, human.** The complete list of sequence motifs found in de novo enrichment analysis of RoDs associated with lower nucleosome density in hESCs compared to differentiated cells. Corresponding E-values and transcription factors associated with similar motifs are indicated.

Protein binding was previously shown to order nucleosomes on a scale larger than the 150 bp observed for most of the RoDs in our analysis[49,50]. We therefore examined how TF binding may affect nucleosome profiles beyond the RoD boundaries in different cell types. To this end we compared the nucleosomal profiles around TF binding motifs in each mouse cell type. Our results show that the average nucleosome occupancy profiles around TF motifs exhibit unique properties depending on the specific TF considered. For the Oct4 motif we observed clear nucleosome phasing emanating away from the site of Oct4 binding in the pluripotent cells but not in somatic TTFs, which lack Oct4 expression (Figure 2.16A). Conversely, for a TF specific for differentiated cells, Hnf4a, we observed phasing in somatic but not pluripotent cells (Figure 2.16B). For a TF that is expressed in ESCs, iPSCs and somatic TTFs, c-Myc/Max, we observed nucleosome phasing in all samples (Figure 2.16C). Interestingly, there is a shift in phasing with c-Myc/Max in pluripotent and somatic cells, which may be indicative of preferential binding of this TF to different genomic regions in these cell types. Together, these data support that local changes in nucleosome occupancy are formed around TF binding sites and suggest that the cell-specific TF expression and binding helps to establish the unique chromatin context for a given cell type[27,51,52].

To further validate that RoDs reflect TF binding sites in mouse we investigated the enrichment of ChIP-Seq signal at these loci, using data for pluripotency-associated TF binding from an independent study[39]. Our results revealed several-fold enrichment of Oct4, Sox2, and Nanog signal at LND RoDs, while no such enrichment was detected for HND RoDs (Figure 2.17). Additionally, the profile of the H3K4me3 histone mark in ESCs, showed a clear drop at the center of LND RoDs, which is consistent with nucleosome depletion. These findings highlight a

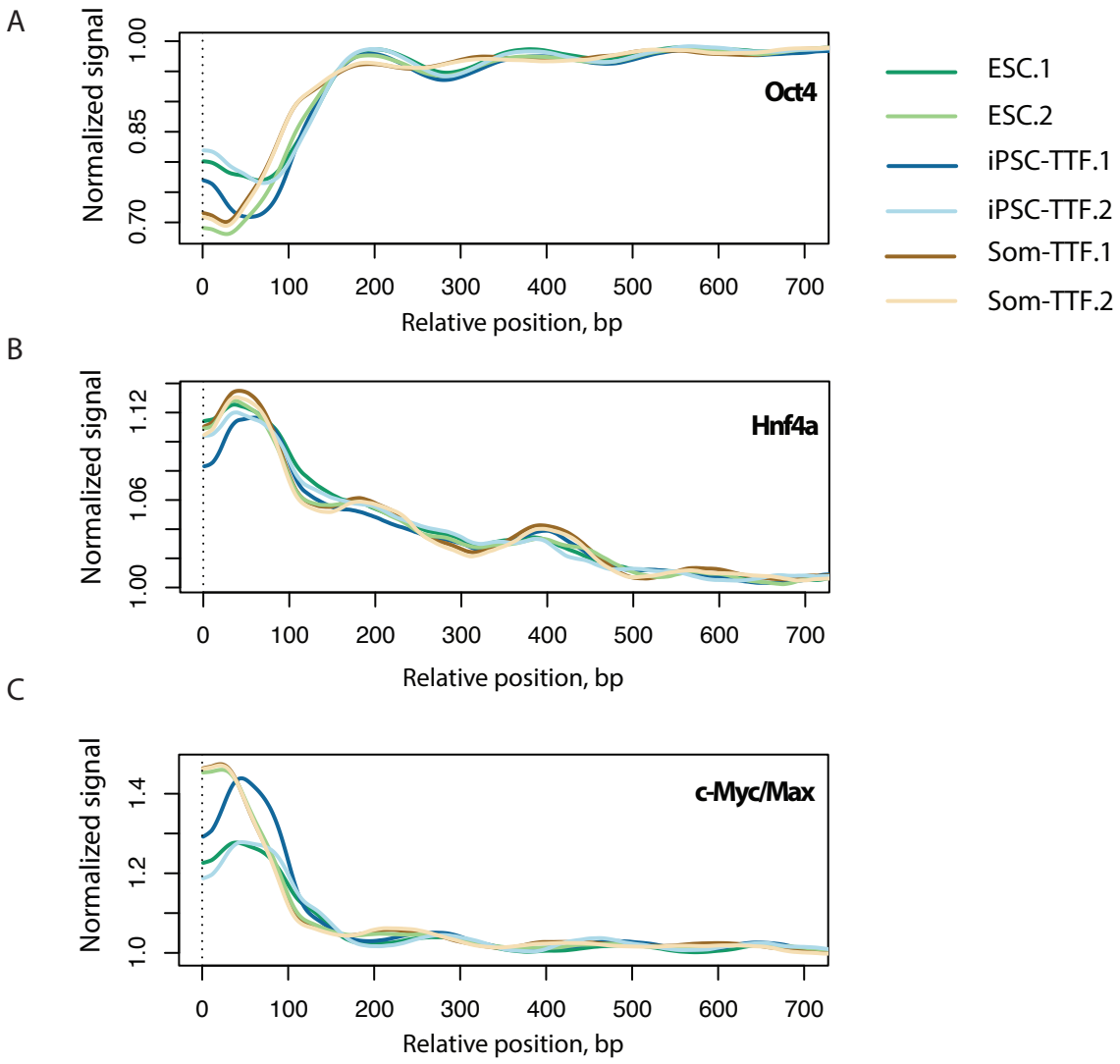



**Figure 2.16 Distribution of nucleosome occupancy around the motifs of selected transcription factors A-C.** Oct4, Hnf4a, and c-Myc/Max, respectively. The occupancy was averaged over all motifs identified in the mouse genome with the selected FDR threshold and the plot was symmetrized relative to the motif center.

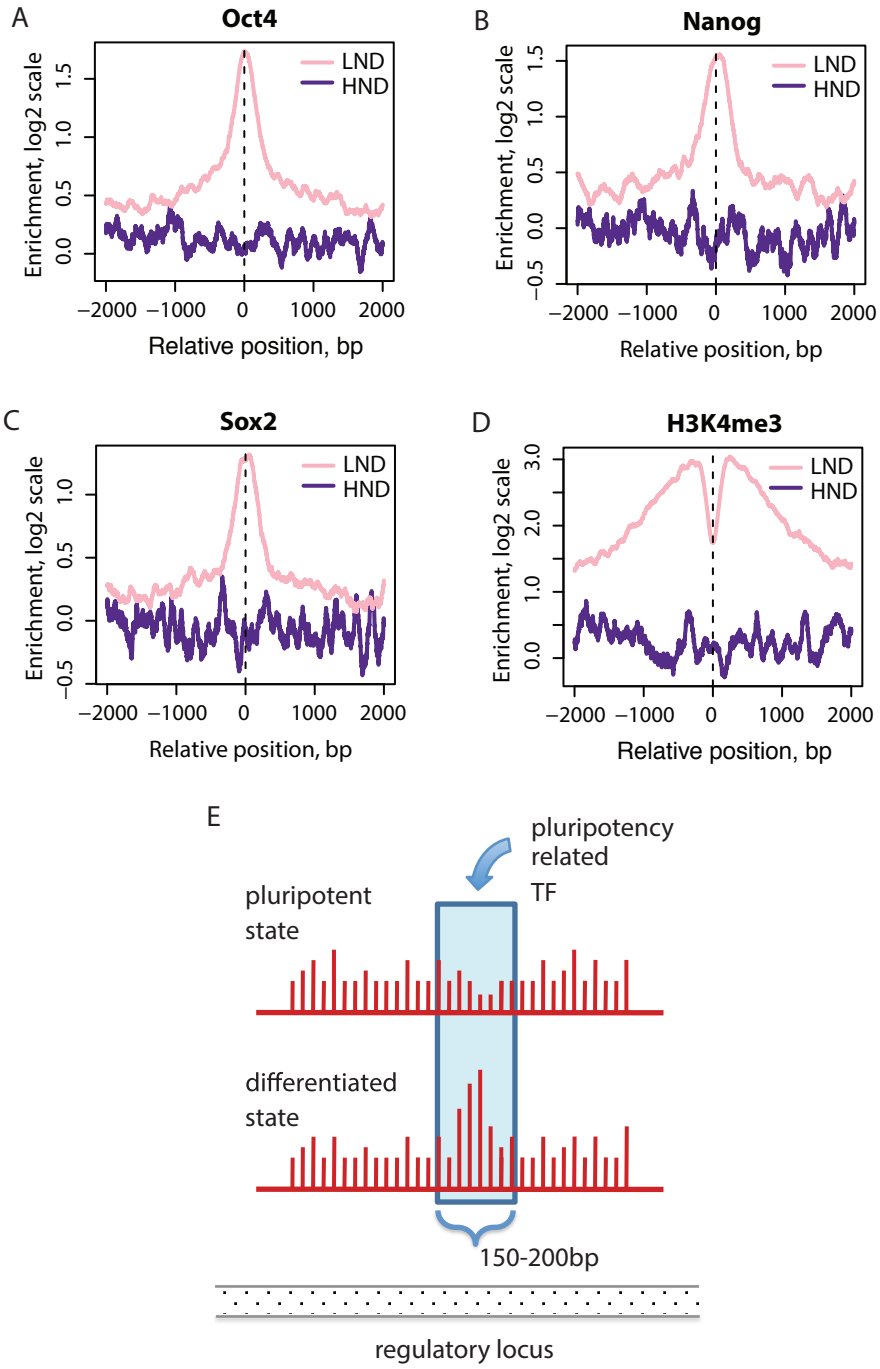



**Figure 2.17 Transcription factor binding at the sites of nucleosome re-arrangement.** **A-D.** Enrichment profiles (ChIP over WCE input) computed in the RoD proximal regions for Oct4, Nanog, Sox2, and H3K4me3 mark. Two classes of RoDs are considered separately, LND (light pink) and HND (purple). **E.** Schematic summary of the observations reported in this chapter. While nucleosome occupancy profiles (red vertical bars) remain similar between the pluripotent and differentiated states, there are punctate regions of significant difference (marked by the light blue rectangle) characterized by lower nucleosome occupancy in the pluripotent state. These regions are predominately the size of a single nucleosome, are enriched in binding motifs of pluripotency-related transcription factors, and occur within regulatory regions, such as gene promoters and enhancers.

possible role for TF binding in the re-arrangement of nucleosomal landscape and suggest that different factors are responsible for emergence of LND and HND RoDs.

Overall, our results revealed that the differences in nucleosome occupancy profiles in pluripotent and somatic cells mostly manifest as punctate changes at the level of individual loci and that these differences tend to cluster at regulatory regions, including developmentally regulated genes and their promoter and enhancer regions. These changes are clustered with genetic genomic elements that control gene expression, indicating their functional importance for determining the regulation of cell status. We conclude that there are not wholesale changes in nucleosome positioning between pluripotent and somatic lineages, but rather specific changes whose location implies a key role in regulating the transition between these cell states.

## Discussion

The primary objective of this study was to determine the nature of changes that occur in nucleosome occupancy profiles upon transition between pluripotent and somatic cells. To address this question we used an MNase digestion assay as the primary experimental tool. We note that while the extent to which MNase-associated bias affects the determination of nucleosome positioning is still debated[53,54], the design of our study, which involves an additional correction step for cleavage bias and focuses on pairwise comparison of the occupancy profiles, minimizes the possibility of artifacts.

One can expect that a dramatic change in cell identity such as that which occurs during somatic cell reprogramming or the differentiation of pluripotent cells would be accompanied by large-scale changes in primary chromatin structure. To our surprise, we detected only a handful of RoDs larger than one kilobase in size. At the same time, we observed several important

features in the re-organization of nucleosomal landscapes associated with pluripotency differentiation and reprogramming. Our main conclusions are that changes in nucleosome occupancy are largely the size of a single-nucleosome, co-localize with pluripotency and differentiation associated protein binding sites, generally have reduced levels of nucleosome occupancy in pluripotent cells compared to somatic cells, and are enriched at enhancers, promoters and within gene regions (Figure 2.11A-C). Comparisons of different classes of regulatory regions revealed that RoDs at enhancers are at least as prevalent as those at TSSs, underscoring the importance of these regions in determining cell state [39,40,55].

Another central conclusion is that fully reprogrammed and characterized iPSCs[29,56] demonstrate nucleosome occupancy patterns similar to those in blastocyst-derived ESCs, with 8-fold fewer RoDs detected between ESCs and iPSCs than between ESCs and differentiated cells in human. Importantly, the nucleosome configuration at enhancers in iPSCs is almost identical to that in ESCs, while it is considerably different from that in fibroblasts. Additionally, the RoDs identified between pluripotent and somatic cells contained binding motifs for key pluripotency factors, suggesting that the nucleosome occupancy changes overlap with genomic regulatory regions that are important for cell identity. Chromatin structure in general, and nucleosome occupancy in particular, could represent an additional and fundamental level of epigenetic memory that must be reset for proper somatic-cell reprogramming in order to allow iPSCs to regain pluripotency and differentiate into an array of cell types[55,57].

Our analysis supports, from a distinct angle, the observation from previous studies that pluripotent cells have chromatin that is more ‘open’ than chromatin in somatic cells. ChIP-seq on H3K9me3 and H3K27me3 suggested that these silencing-associated marks cover over three

times more of the genome in differentiated cells when compared to ESCs[58]. In addition, the nuclei of pluripotent cells have macroscopic characteristics of less-condensed chromatin, and histone turnover appears more dynamic in pluripotent cells[4]. Here, we observe that a majority of the detected RoDs are associated with lower nucleosome occupancy in pluripotent cells when compared to somatic cells and, furthermore, that pluripotent-associated enhancers have low nucleosome density on a kilobase scale in pluripotent cells versus differentiated. The lower levels of nucleosome occupancy in pluripotent cells correlates with function, since it predominantly occurs at active chromatin regions, including ESC-specific enhancers. Thus we conclude that the more permissive chromatin configuration in pluripotent cells is enabled not only through reduction of the repressive chromatin but also through local changes in the nucleosomal landscapes in euchromatic regions.

While most of RoDs are of the size of a single nucleosome, we note that protein binding may induce larger-scale rearrangement of chromatin, such as the increased nucleosome phasing observed in Figure 2.16. However, much deeper sequencing and a larger number of replicates would be required to identify a 'complete' set of RoDs which would include such changes at individual loci. In combination with protein-binding motif information, our current approach can be used for simultaneous identification of nucleosome re-arrangement and differential binding for a range of transcription factors in one assay, when such data are available. This approach could be further enhanced by analyzing the digested fragments of sub-nucleosomal sizes and/or by using multiple levels of digestion for the same sample to preferentially profile genomic regions of different accessibility[31,59]. Such a comprehensive

approach would help us better understand how changes in chromatin organization translate into changes in gene expression and cell identity.

## Methods

### Experimental Procedures

Cell Culture

Human ESCs and iPSCs were maintained on Geltrex (Life Technologies) in mTeSR1 (Stem Cell Technologies). H1-OGN ESCs[60] and iPSCs[29] were a gift from George Daley and were functionally characterized previously[29,60]. These cells exhibited the expected *in vitro* molecular and functional properties of human pluripotent cells in our hands, but showed low to no OCT4-GFP reporter expression. Experiments were carried out with H1-OGN ESCs between passage 76-77 and iPSCs between passage 14-17. Differentiated fibroblasts were made from H1-OGN ESCs and were used between passages 7-14.

Mouse ESCs and iPSCs were maintained on MEF feeder layers (Specialty Media) in DMEM containing 15% heat-inactivated fetal bovine serum (Hyclone) supplemented with 1000U/mL LIF (Chemicon). The following mouse cell lines were used in these studies: A5 ESCs (ESC.1), A6 ESCs (ESC.2), A4 iPSCs (iPS.TTF.1), A5 iPSCs (iPS.TTF.2), and Hep2 iPSCs (iPS.Liver). All isogenic lines were created from mice containing the stable integration of doxycycline (dox)-inducible reprogramming factors (Oct4, Sox2, Klf4, and c-Myc), and therefore do not vary with regard to viral integration sites. All experiments were initiated with cell lines between passage 15 and 22. Primary TTFs and liver were obtained as secondary derivates from B6/129 neonatal mice aged between 7 to 14 days postpartum. These mice and cell lines have been functionally characterized and were previously reported[33].

Chromatin digestion with MNase

Human cells were expanded to approximately $1x10^8$ cells and cross-linked with 1.1% formaldehyde for 10 minutes at room temperature. Nuclei were isolated and treated with a range of four MNase concentrations for 15 minutes at room temperature. Cross-link reversal was performed at 65$^{o}$C for at least 16 hours followed by an RNase and subsequent proteinase K digestion. DNA was purified by phenol-chloroform extraction. Ampure SPRI beads (Beckman Coulter) were used in a double size selection with ratios of 0.7X and 1.7X to obtain a range of fragment sizes from approximately 100 bp to 300 bp. The resulting sample contains a majority of mono-nucleosomal fragments with some smaller and di-nucleosome-sized fragments with high reproducibility. The resulting fragments from each MNase concentration in the range were prepared individually for barcoded sequencing on an Illumina HiSeq instrument. Mapped read from all concentration were later pooled for analysis.

Each murine cell type was expanded to approximately $3 \times 10^7$ cells and then pretreated with mild detergents (0.2% Tween-20 and 0.2% Triton X-100) for 5 minutes followed by a 1.1% formaldehyde treatment for 10 minutes to preserve chromatin structure. Nuclei were then prepared from the crosslinked cells and the chromatin treated with a range of micrococcal nuclease (MNase) concentrations for 15 minutes at room temperature. A range of digestion conditions was employed to sample both hyper- and hypo-accessible chromatin regions to MNase digestion. Cross-links were then reversed for 16 hours at 55$^{o}$C along with proteinase K digestion and DNA harvested via phenol-chloroform. Samples were then run on 1% agarose gels and the resulting mononucleosomal DNA fragments (approximately 150bp) were gel purified, pooled, and prepared for sequencing on an Illumina HiSeq instrument.

Illumina HiSeq Library preparation and sequencing

100 ng of mononucleosome DNA was used for library preparation, with limited numbers of PCR amplification rounds[61], and genomic alignments of paired-end 50 bp reads were performed using Bowtie[62] followed by further tag processing and filtering with the SPP workflow[29]. All alignments and annotations used the human genome assembly hg19 and the mouse genome assembly mm9.

## Bioinformatic and statistical data analysis

Sequencing data preprocessing and initial analysis

Sequenced 50bp paired-end tags were mapped to the human genome (hg19) or mouse (mm9) for the corresponding cell types using the Bowtie aligner v. 0.12.7[62]. Only uniquely mapped tags with no more than two mismatches in the first 28 bp of the tag were retained. Genomic positions with the numbers of mapped tags above the significance threshold of z-score=7 were identified as anomalous, and the tags mapped to such positions were discarded. The coordinates of the genes were taken according to the annotations for hg19 and mm9 versions of the human and mouse genomes respectively. Gene ontology analysis was performed using the Gene Ontology Term Finder web-server (http://go.princeton.edu/cgi-bin/GOTermFinder/GOTermFinder)[63]. The gene proximal profiles were calculated and plotted as described previously[30,64].

GC-content normalization

The correction coefficient for each read was computed in such a way that the resulting genome-wide distributions of GC-content become similar to the target GC-content distribution (Gaussian distribution with mean GC=50% and 48% and variance=7.5%). Specifically, all reads were stratified according to the GC-content of the regions +/-100bp around the pair-end read centers and the correction coefficients were computed as ratios of the histograms corresponding to experimental and theoretical GC-content distributions with 1% GC content step. The value of GC=50% was used to obtain main results in the study.

Identification of regions of difference in nucleosome occupancy

The P-values of difference were estimated for frequency of reads summarized within 150-bp non-overlapping bins. The P-value calculation was based on the negative binomial distribution, with variance and mean estimated based on the replicate profiles produced for each cell type, as implemented in R package DESeq[65]. Default parameters of DESeq package were used for computations. To account for local context of nucleosome occupancy, the estimation of significance of the nucleosome occupancy changes within bins was performed independently in 25 Kb segments with a 12.5 Kb step, hence generating two significance values for each bin. The more conservative estimate was retained for further analysis. The bins exhibiting significant changes in nucleosome occupancy between the samples separated by less than 100 bp were merged to form regions of difference.

Estimation of statistical significance

Significance estimations were performed using R (http://www.r-project.org). Abundances of RoDs in genomic regions were compared to the corresponding values obtained for the randomized RoD distributions using non-parametric Wilcoxon test (as implemented in function "wilcox.test" from the package "stats"). Only the regions of the genome that had non-zero tag counts were used in randomization (at least 1000 randomizations were performed in each case).

Motif analysis

Motif analysis was performed using web-base service MEME-ChIP[66]. Motifs at least six base pairs in length identified with E-value threshold of 1e-5 were reported. Both palindromic and non-palindromic motifs were allowed. The motifs found in the test sequences were matched against JASPAR (CORE-2009) or UniPROBE (mouse) databases to identify similarity with known protein motifs using tools implemented in MEME-ChIP with default parameters.

Data availability

Data sets are deposited in the NIH GEO database under Series GSE59064.

## Acknowledgments

We thank S. Bowman and M. Simon for optimizing sequencing library preparation, Z. Wang, C. Woo, J. Dennis, and the Kingston and Park labs for helpful discussions, G.Q. Daley for human cell lines, and the MGH Molecular Biology NextGen Sequencing Core. J.A.W., R.E.K., and P.J.P were supported by the National Institute Of General Medical Sciences of the NIH

(F32GM093491 to J.A.W; R01GM043901 and R37GM048405 to R.E.K.; R01GM082798 to P.J.P).

K.H. was supported by the NIH grants R01HD058013 and DP2OD003266.

## Contributions

A.C. performed human cell line experiments, J.A.W. performed mouse cell line experiments, B.A. helped develop bioinformatic tools, M.S. and K.H. isolated all mouse cell lines and functionally characterized the mouse pluripotent lines, A.D. provided expertise in interpretation of the results, M.Y.T. analyzed the data, A.C., J.A.W., M.Y.T., P.J.P. and R.E.K. designed the study, interpreted the results and wrote the paper. All authors read and contributed editing to the manuscript during its preparation.

**Competing financial interests**

The authors declare no competing financial interests.

## References

1 Gao, X. *et al.* ES cell pluripotency and germ-layer formation require the SWI/SNF chromatin remodeling component BAF250a. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 6656-6661 (2008).

2 Ho, L. *et al.* An embryonic stem cell chromatin remodeling complex, esBAF, is essential for embryonic stem cell self-renewal and pluripotency. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 5181-5186 (2009).

3 Loh, Y. H., Zhang, W., Chen, X., George, J. & Ng, H. H. Jmjd1a and Jmjd2c histone H3 Lys 9 demethylases regulate self-renewal in embryonic stem cells. *Genes & development* **21**, 2545-2557 (2007).

4 Meshorer, E. *et al.* Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Dev Cell* **10**, 105-116 (2006).

5 Fussner, E. *et al.* Constitutive heterochromatin reorganization during somatic cell reprogramming. *The EMBO journal* **30**, 1778-1789, doi:10.1038/emboj.2011.96 (2011).

6 Gaspar-Maia, A., Alajem, A., Meshorer, E. & Ramalho-Santos, M. Open chromatin in pluripotency and reprogramming. *Nature reviews. Molecular cell biology* **12**, 36-47, doi:10.1038/nrm3036 (2011).

7 Wang, J. *et al.* A protein interaction network for pluripotency of embryonic stem cells. *Nature* **444**, 364-368 (2006).

8 Yildirim, O. *et al.* Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. *Cell* **147**, 1498-1510, doi:10.1016/j.cell.2011.11.054 (2011).

9 Sansam, C. G. & Roberts, C. W. Epigenetics and cancer: altered chromatin remodeling via Snf5 loss leads to aberrant cell cycle regulation. *Cell Cycle* **5**, 621-624 (2006).

10 Reisman, D., Glaros, S. & Thompson, E. A. The SWI/SNF complex and cancer. *Oncogene* **28**, 1653-1668 (2009).

11 Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184**, 868-871 (1974).

12 Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature* **389**, 251-260, doi:10.1038/38444 (1997).

13 Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128**, 707-719 (2007).

14 Morse, R. H. Transcription factor access to promoter elements. *J Cell Biochem* **102**, 560-570 (2007).

15 Dennis, J. H. *et al.* Independent and complementary methods for large-scale structural analysis of mammalian chromatin. *Genome research* **17**, 928-939 (2007).

16 Li, Z., Schug, J., Tuteja, G., White, P. & Kaestner, K. H. The nucleosome map of the mammalian liver. *Nature structural & molecular biology* **18**, 742-746, doi:10.1038/nsmb.2060 (2011).

17 Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887-898 (2008).

18 Teif, V. B. *et al.* Genome-wide nucleosome positioning during embryonic stem cell development. *Nature structural & molecular biology* **19**, 1185-1192, doi:10.1038/nsmb.2419 (2012).

19 Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516-520, doi:10.1038/nature10002 (2011).

20 Woo, C. J., Kharchenko, P. V., Daheron, L., Park, P. J. & Kingston, R. E. A Region of the Human HOXD Cluster that Confers Polycomb-Group Responsiveness. *Cell* **140**, 99-110 (2010).

21 Li, Z. *et al.* Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell* **151**, 1608-1616, doi:S0092-8674(12)01403-1 [pii] 10.1016/j.cell.2012.11.018 (2012).

22 Tolstorukov, M. Y. *et al.* Swi/Snf chromatin remodeling/tumor suppressor complex establishes nucleosome occupancy at target promoters. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 10165-10170, doi:10.1073/pnas.1302209110 (2013).

23 Weiner, A., Hughes, A., Yassour, M., Rando, O. J. & Friedman, N. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome research* **20**, 90-100, doi:gr.098509.109 [pii]10.1101/gr.098509.109 (2010).

24 Henikoff, S., Henikoff, J. G., Sakai, A., Loeb, G. B. & Ahmad, K. Genome-wide profiling of salt fractions maps physical properties of chromatin. *Genome research* **19**, 460-469, doi:gr.087619.108 [pii]10.1101/gr.087619.108 (2009).

25 Zwaka, T. & Thomson, J. Homologous recombination in human embryonic stem cells. *Nature biotechnology* **21**, 319-321, doi:10.1038/nbt788 (2003).

26 Park, I.-H. *et al.* Reprogramming of human somatic cells to pluripotency with defined factors. *Nature* **451**, 141-146, doi:10.1038/nature06534 (2008).

27 Yuan, G. *et al.* Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* **309**, 626-630, doi:10.1126/science.1112178 (2005).

28 Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442**, 772-778 (2006).

29 Kharchenko, P. V., Woo, C. J., Tolstorukov, M. Y., Kingston, R. E. & Park, P. J. Nucleosome positioning in human HOX gene clusters. *Genome Res* **18**, 1554-1561 (2008).

30 Tolstorukov, M. Y. *et al.* Histone variant H2A.Bbd is associated with active transcription and mRNA processing in human cells. *Molecular cell* **47**, 596-607, doi:10.1016/j.molcel.2012.06.011 (2012).

31 Henikoff, J. G., Belsky, J. A., Krassovsky, K., MacAlpine, D. M. & Henikoff, S. Epigenome characterization at single base-pair resolution. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 18318-18323, doi:10.1073/pnas.1110731108 (2011).

32 Koerber, R. T., Rhee, H. S., Jiang, C. & Pugh, B. F. Interaction of transcriptional regulators with specific nucleosomes across the Saccharomyces genome. *Molecular cell* **35**, 889-902, doi:10.1016/j.molcel.2009.09.011 (2009).

33 Stadtfeld, M., Maherali, N., Borkent, M. & Hochedlinger, K. A reprogrammable mouse strain from gene-targeted embryonic stem cells. *Nature methods* **7**, 53-55, doi:10.1038/nmeth.1409 (2010).

34 Gaffney, D. J. *et al.* Controls of nucleosome positioning in the human genome. *PLoS genetics* **8**, e1003036, doi:10.1371/journal.pgen.1003036 (2012).

35 Chung, H.-R. *et al.* The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS One* **5**, doi:10.1371/journal.pone.0015754 (2010).

36 Johnson, W. E. *et al.* Model-based analysis of tiling-arrays for ChIP-chip. *Proceedings of the National Academy of Sciences of the United States of America* **103**, 12457-12462, doi:0601180103 [pii] 10.1073/pnas.0601180103 (2006).

37 Cheung, M. S., Down, T. A., Latorre, I. & Ahringer, J. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic acids research* **39**, e103, doi:gkr425 [pii] 10.1093/nar/gkr425 (2011).

38 Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic acids research* **40**, e72, doi:gks001 [pii] 10.1093/nar/gks001 (2012).

39 Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153**, 307-319, doi:10.1016/j.cell.2013.03.035 (2013).

40 Hnisz, D. *et al.* Super-Enhancers in the Control of Cell Identity and Disease. *Cell*, doi:S0092-8674(13)01227-0 [pii] 10.1016/j.cell.2013.09.053 (2013).

41 Chen, K. *et al.* DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome research* **23**, 341-351, doi:gr.142067.112 [pii] 10.1101/gr.142067.112 (2013).

42 Bilic, J. & Izpisua Belmonte, J. C. Concise review: Induced pluripotent stem cells versus embryonic stem cells: close enough or yet too far apart? *Stem cells (Dayton, Ohio)* **30**, 33-41, doi:10.1002/stem.700 (2011).

43 Bilic, J. & Izpisua Belmonte, J. C. Concise review: Induced pluripotent stem cells versus embryonic stem cells: close enough or yet too far apart? *Stem cells* **30**, 33-41, doi:10.1002/stem.700 (2012).

44 Papp, B. & Plath, K. Epigenetics of reprogramming to induced pluripotency. *Cell* **152**, 1324-1343, doi:S0092-8674(13)00277-8 [pii] 10.1016/j.cell.2013.02.043 (2013).

45 Loh, Y. H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat Genet* **38**, 431-440 (2006).

46 Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-1117, doi:S0092-8674(08)00617-X [pii] 10.1016/j.cell.2008.04.043 (2008).

47 Chambers, I. & Tomlinson, S. R. The transcriptional foundation of pluripotency. *Development* **136**, 2311-2322, doi:136/14/2311 [pii] 10.1242/dev.024398 (2009).

48 Matsuda, T. *et al.* STAT3 activation is sufficient to maintain an undifferentiated state of mouse embryonic stem cells. *The EMBO journal* **18**, 4261-4269, doi:10.1093/emboj/18.15.4261 (1999).

49 Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS genetics* **4**, e1000138, doi:10.1371/journal.pgen.1000138 (2008).

50 Hu, G. *et al.* Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome research* **21**, 1650-1658, doi:10.1101/gr.121145.111 (2011).

51 Kornberg, R. D. & Stryer, L. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic acids research* **16**, 6677-6690 (1988).

52 Mavrich, T. *et al.* Nucleosome organization in the Drosophila genome. *Nature* **453**, 358-362, doi:10.1038/nature06929 (2008).

53 Chung, H. R. *et al.* The effect of micrococcal nuclease digestion on nucleosome positioning data. *PloS one* **5**, e15754, doi:10.1371/journal.pone.0015754 (2010).

54 Allan, J., Fraser, R. M., Owen-Hughes, T. & Keszenman-Pereyra, D. Micrococcal nuclease does not substantially bias nucleosome mapping. *Journal of molecular biology* **417**, 152-164, doi:10.1016/j.jmb.2012.01.043 (2012).

55 Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* **151**, 994-1004, doi:S0092-8674(12)01298-6 [pii] 10.1016/j.cell.2012.09.045 (2012).

56 Stadtfeld, M. *et al.* Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. *Nature* **465**, 175-181, doi:10.1038/nature09017 (2010).

57 Soufi, A. & Zaret, K. S. Understanding impediments to cellular conversion to pluripotency by assessing the earliest events in ectopic transcription factor binding to the genome. *Cell Cycle* **12**, 1487-1491, doi:24663 [pii] 10.4161/cc.24663 (2013).

58 Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell stem cell* **6**, 479-491, doi:10.1016/j.stem.2010.03.018 (2010).

59 Bryant, G. O. *et al.* Activator control of nucleosome occupancy in activation and repression of transcription. *PLoS biology* **6**, 2928-2939, doi:10.1371/journal.pbio.0060317 (2008).

60 Zwaka, T. P. & Thomson, J. A. Homologous recombination in human embryonic stem cells. *Nature biotechnology* **21**, 319-321, doi:10.1038/nbt788 nbt788 [pii] (2003).

61 Bowman, S. K. *et al.* Multiplexed Illumina sequencing libraries from picogram quantities of DNA. *BMC Genomics* **14**, 466, doi:1471-2164-14-466 [pii] 10.1186/1471-2164-14-466 (2013).

62 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25, doi:gb-2009-10-3-r25 [pii] 10.1186/gb-2009-10-3-r25 (2009).

63 Boyle, E. I. *et al.* GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. *Bioinformatics* **20**, 3710-3715, doi:10.1093/bioinformatics/bth456 bth456 [pii] (2004).

64 Tolstorukov, M. Y., Kharchenko, P. V., Goldman, J. A., Kingston, R. E. & Park, P. J. Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. *Genome research* **19**, 967-977, doi:gr.084830.108 [pii] 10.1101/gr.084830.108 (2009).

65 Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol* **11**, R106, doi:gb-2010-11-10-r106 [pii] 10.1186/gb-2010-11-10-r106 (2010).

66 Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27**, 1696-1697, doi:btr189 [pii] 10.1093/bioinformatics/btr189 (2011).

# Chapter 3 Using MNase titrations to probe chromatin accessibility

## Contributions

This chapter, modified from a draft manuscript to reflect the work of April Cook, reflects the work of several additional people. April Cook performed the human K562 cell experiments, Sarah Bowman performed *Drosophila melanogaster* S2 cell experiments, Michael Tolstorukov and April Cook analyzed the data, A.C., S.B., M.Y.T., and Bob Kingston designed the study, interpreted the results, and are writing the paper.

## Abstract

The structure of chromatin controls the access of regulatory factors to DNA. In basic form nucleosomes on DNA in chromatin are often described as 'beads on a string'. However, chromatin exists at a wide range of compaction levels depending on the regulatory state of the genomic region in question. Micrococcal nuclease (MNase), an enzyme that digests linker DNA between nucleosomes, has long been used to map nucleosome occupancy. However, current methodology maps only the nucleosomes accessible at a limited range of MNase digestion. Here we present a new methodology for characterizing a broad range of levels of accessibility through the use of an MNase titration series. We find that a range of four MNase digestion levels produces four different nucleosome occupancy maps. Different extents of digestion have an impact on how MNase-seq data correlates with chromatin features including histone tail modifications, transcription factors, and remodelers. Further, considering these maps together, we are able to assign a chromatin state, 'open' or 'closed', to a region based on the pattern of occupancy changes seen at a particular locus. These open and closed regions correlate with

marks thought to be active or inactive. In addition to grouping open and closed states we develop a new metric, called MACC, that is a quantitative measure of the degree of change between states. We find that the degree of openness as quantified by MACC both upstream and downstream of the transcription start site correlates with transcription level. Additionally, MNase titration with MACC calculation can be used to probe a wide range of chromatin states and assess both global and local chromatin accessibility.

## Introduction

Each cell's genome is the blueprint used for production of all the building blocks that give that cell an identity and allow it to perform cellular functions. To use the information in the genome, a fundamentally important process, transcription, occurs not on naked DNA, but on a chromatin template. RNA Polymerase II (PolII) must access the DNA to transcribe RNA that will then go on to be translated into protein or perform a regulatory role.

Chromatin is a heterogeneous mix of nucleic acid and proteins, all packed in a nucleus. The basic repeating structural unit of chromatin is the nucleosome: a histone octamer wrapped in approximately 150 base pairs of DNA. Nucleosomes organize the vast majority of the genome and they control access to DNA in a context-dependent manner[1]. Because of the critical role of chromatin in regulating access to DNA, there has been an increasing interest in characterizing and understanding chromatin architecture. There have been a proliferation of covalent histone tail modification maps[2,3], studies on chromatin protein structure and function[4,5], and studies determining organization of chromatin features in the nucleus using microscopy[6,7]. While our increasing understanding of the above areas has helped us understand the role of chromatin in

the cell, these studies leave open questions about a fundamental regulatory role of chromatin, which is where it is providing physical access to the DNA.

The physical properties of nucleosomes, in particular, stability, vary throughout the genome. Histones in nucleosomes are subject to covalent modifications and replacement with variant proteins, which can affect the physical properties of nucleosomes and their ability to form higher order structures[8,9]. Also, the flexibility of underlying DNA is sequence-dependent, which results in nucleosome stability being partially dependent on its genomic location[10,11]. Thus, mapping the genomic location of nucleosomes will enable better insight into the functional organization of the genome and provide a reference for better understanding how the epigenome contributes to cellular processes.

DNA accessibility and nucleosome positioning are often measured with nuclease assays. Common approaches include digestion of chromatin with MNase or DNaseI nucleases followed by high-throughput analysis of the digestion products[12-14]. When considered together with data from genome-wide chromatin immunoprecipitation studies, these assays can be used for profiling additional chromatin properties, including nucleosome turnover or the structural state of regions with particular histone variants present[12,15]. Also, a methodology using MNase along with varying salt concentrations to probe nucleosome stability was recently used for the analysis of fly chromatin[16]. The use of nucleases to probe chromatin is a trusted method; here we aim to improve upon these methods.

Digestion assays have been reported to have to have intrinsic biases[17]. Micrococcal nuclease (MNase), has a sequence preference[18,19] and its use in profiling assays is sensitive to the level of enzyme activity. Minor differences in digestion conditions can have a profound

effect on experimental outcome. For example, it has been reported that different concentrations of MNase produce different nucleosome occupancy profiles at local regions of the genome. These differences can manifest as a region seemingly covered by nucleosomes in one condition but depleted in another[20-22]. Biases in digestion assays can result in poor reproducibility and hinder sample-to-sample comparison. Studies using MNase in mammalian systems where chromatin remodeling has been well-characterized have reported finding only minor differences between conditions[15,23], raising the question of whether changing experimental conditions may improve the resolution of these studies. Experimental and computational approaches have been proposed to deal with these biases that attempt to 'standardize' the nucleosome occupancy profiles through optimizing digestion conditions and/or normalizing the generated data, including those in the previous chapter [24], West, in press. As with any technique, eliminating noise and biases and addressing the optimal use of the tool before data analysis is ideal.

Here we describe a novel approach to study the physical organization of chromatin; specifically, we seek to measure how accessible every nucleosome in the genome is. A characterization of the relative accessibility of every nucleosome to MNase is important for understanding the activity of regulatory factors. This approach leverages the power of MNase digestion of chromatin with a range of digestion levels rather than attempting to correct for the effects of variability of extent of digestion in one condition. We use several MNase concentrations (MNase titration) to create independent datasets corresponding to different digestion levels. To process and analyze these data we developed a specialized bioinformatic methodology; this methodology was used to produce accessibility maps for both human and

drosophila genomes. We demonstrate that accessibility is a useful measure of functional chromatin. We propose that measuring accessibility through the use of MNase titrations is a biologically relevant way of probing chromatin features across the genome, including regulatory loci. We produce a new metric that allows quantitative measurement and comparison of accessibility that is broadly applicable.

## Results

To obtain a comprehensive picture of the role of nucleosome placement in regulation of genomic DNA accessibility we digested chromatin from human K562 cells using series different concentrations of micrococcal nuclease (MNase titrations). To obtain nucleosome-protected DNA, cross-linked cells were treated with a 0.4-fold titration series of MNase (18 total concentrations) and the outcome was monitored by agarose gel electrophoresis. From this set a range of four enzyme concentrations (5.4, 20.6, 79.2, and 304 U/mL) was chosen such that the lowest concentration produced a minimal number of mononucleosomal fragments, and the highest concentration produced predominantly mononucleosomal (Figure 3.1A). A four-fold digestion series was also performed in *D. melanogaster* S2 cells (1.5, 6.25, 25 and 100 U/µL)(Figure 3.1B). Following a low-stringency size selection to remove most DNA over 1000 bp (see Methods), the remaining digestion products were subjected to library construction and paired-end sequencing.

Sequence quality can impact genome wide analysis. We had total tag counts of at least 25 million per a digestion point and the libraries showed high complexity with low percentages of duplicate fragments. The average fragment length shows some dependence on the degree of digestion, with the mononucleosomal fragment length ranging from 150-175 base pairs with an

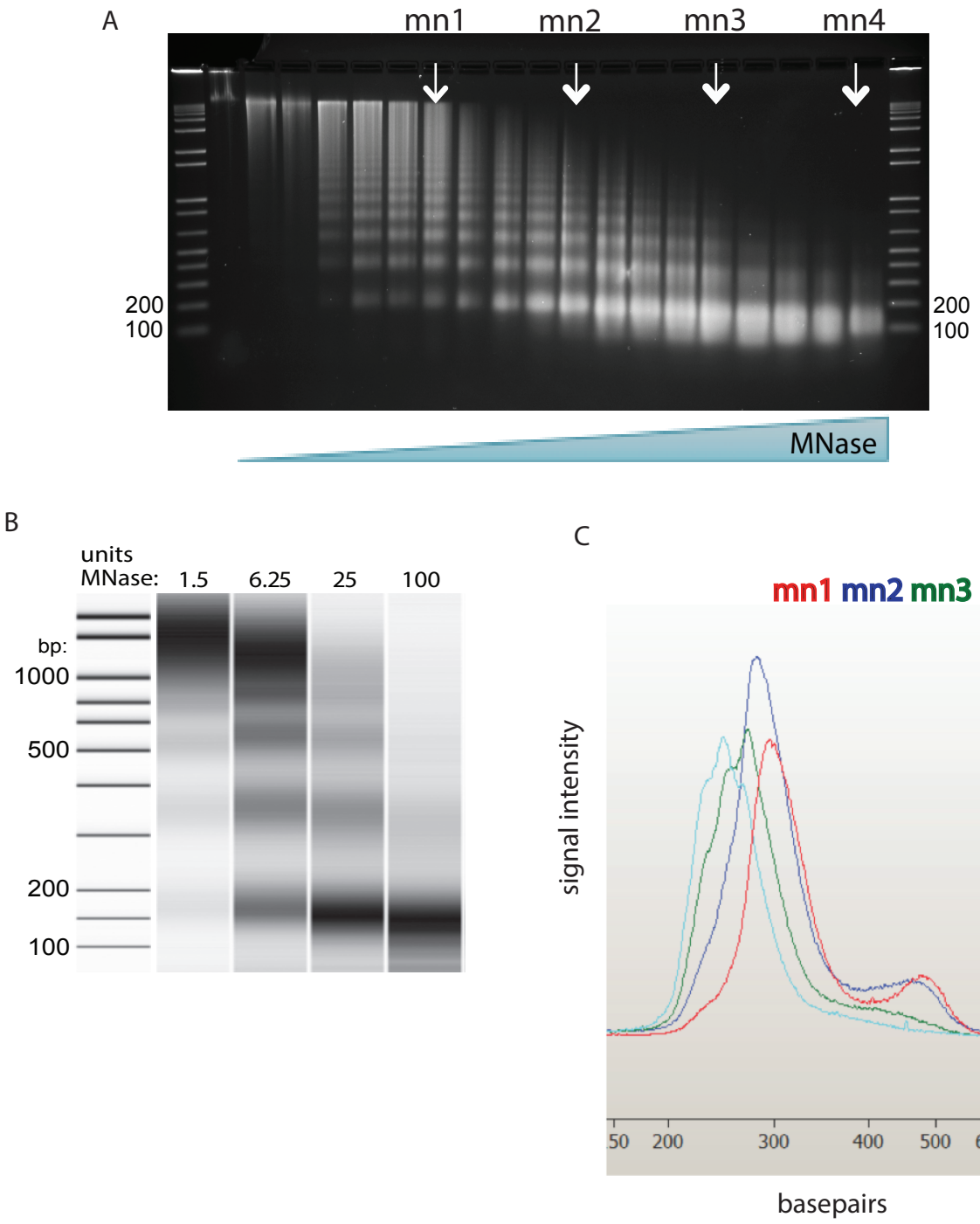



**Figure 3.1 Characteristics of MNase-digested DNA. A.** 1% agarose gel with resulting K562 DNA ladders following eighteen 0.4 fold titrations of MNase. Lanes 1 and 21 are 1kb+ ladders (100 and 200 bp are marked), lane 2 is DNA from an undigested control. Lanes 3-20 are x-304 U/mL. **B.** Bioanalyzer capillary electrophoresis output for S2 cells. Lane1 is a ladder and units of Mnase/ul are marked in the remaining sample wells. **C.** Bioanalyzer electropherogram of the four K562 titration sample libraries sequenced in this study, MNase1 (red) was digested with low MNase, and MN4 (light blue) with the highest MNase. The x axis is DNA size in basepairs.

abundance of sub-nucleosome sized fragments in the sample with the highest concentration of MNase used (Figure 3.1C). It should also be noted that the samples used were cross-linked and a fraction of DNA fragments could have been protected by the DNA-binding factors of a non-histone nature. Since the current study is devoted to the analysis of DNA accessibility rather than nucleosome positioning *per se*, we conclude that this factor should not have a considerable effect on downstream analyses.

## General trends observed

A characteristic feature of nucleosomal landscapes is a stereotypical pattern of nucleosome occupancy around transcription start sites. This pattern of occupancy is a useful point of comparison when evaluating the relationship between different datasets. We assessed the average nucleosome occupancy for 'expressed' and 'silent' genes at TSSs for each digestion condition (Figure 3.2A,B). This analysis revealed strong dependence of the shape of the aggregate TSS-proximal profiles on the degree of the chromatin digestion. For instance, the position of nucleosome "-1" immediately upstream of TSS can be detected only under the light digestion conditions, while observed nucleosome occupancy inside gene bodies is higher under deep digestion conditions. The same general pattern was seen in *D. melanogaster* data (Figure 3.2C,D). We note that the average profiles (red lines in Figure 3.2C,D), generated by combining all the sequenced reads produced for different digestion points in these samples would mask these differences between the digestion points and have a shape similar to those reported in previous publications[12,13,23,25-27].

Typically, a single MNase digestion condition is used in MNase-produced nucleosome occupancy maps. There is no set protocol for determining this condition. In the literature, this

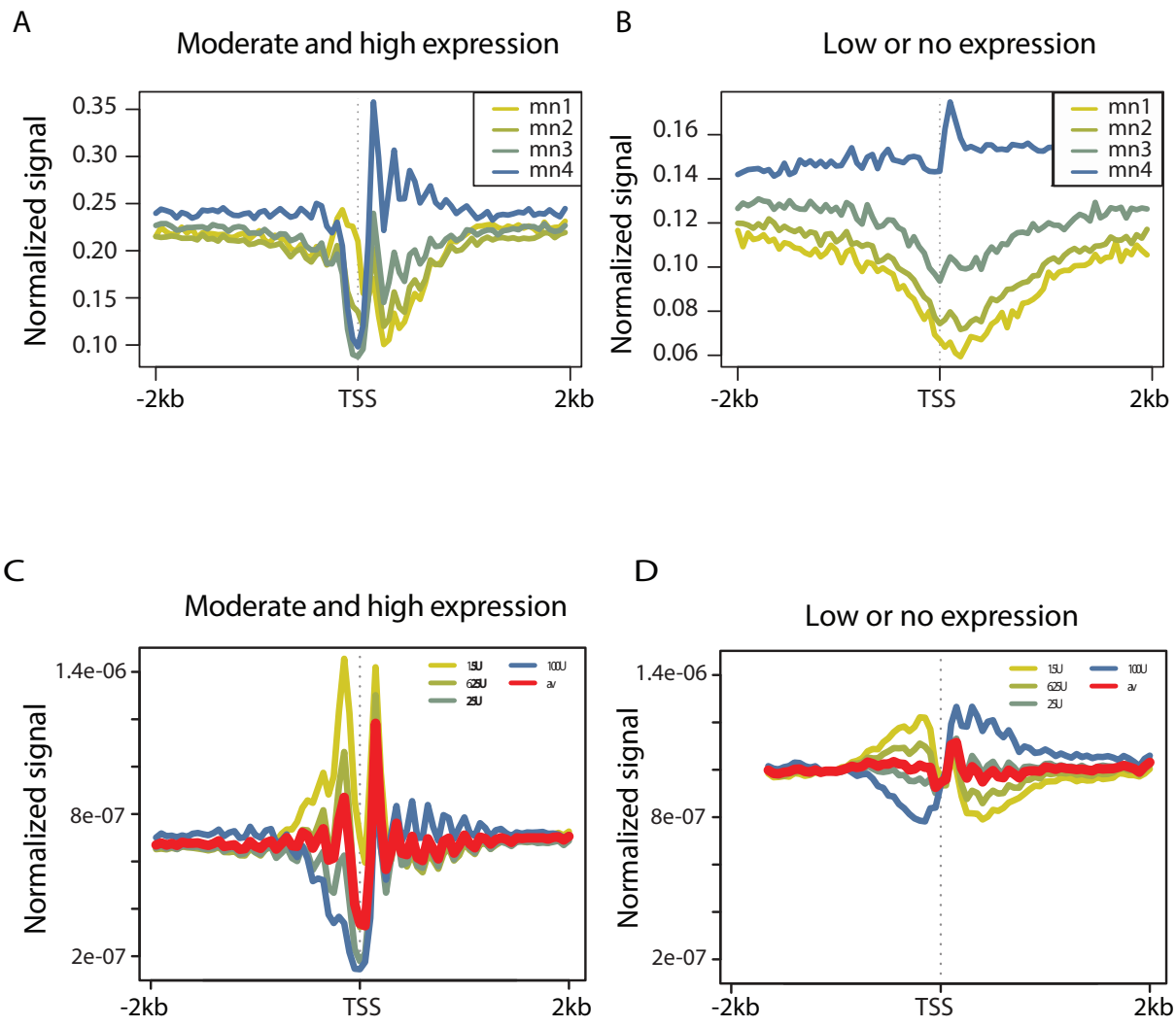



**Figure 3.2 MNase concentration determines nucleosome occupancy profile.** **A**. Average nucleosome occupancy profiles generated around human K562 TSS for silent (left) and expressed (right) genes. Different colors indicate MNase concentrations used for the digestion (5.4, 20.6 79.2 and 304 U/mL), with dark blue color corresponding to the highest and bright yellow to the lowest concentrations. **B**. As in A, but with D. melanogaster S2 cell data (1.5, 6.25, 25 and 100 U/uL MNase). The red line on each plot represents the profile obtained for 'pooled' data based on all four concentrations.

single condition yields anywhere from one single mono-nucleosome sized band to a ladder of multi-nucleosomal bands when run on a gel, and only the mononucleosome-sized band is used for analysis. These studies seem to operate under the assumption that chromatin digested to any extent that produces a mono nucleosomal band can be used to map all nucleosomes along the genome. Said differently, the assumption is that a range of extents of digestion would always produce the only nucleosome map possible, and digestion with a MNase titration series should produce a series of identical maps showing the true occupancy of nucleosomes across the genome. What we find, however, is that DNA loci can have different levels of accessibility to MNase. Our interpretation of this is that there are a variety of scenarios impacting chromatin response to MNase probing. Under light digestion conditions more accessible regions are preferentially profiled. These accessible regions get digested to fragments of sub-nucleosomal sizes under deeper digestion conditions, and eventually become so small that they are lost during library preparation. In contrast, regions of low accessibility do not produce strong signal under light digestion conditions, since higher MNase activity is required to digest such loci into mononucleosomal fragments. These scenarios are further illustrated in Figure 3.3A, arrows show tag frequencies at loci responding differently to MNase titration.

### MNase titration data at known features

Nucleosome occupancy at the TSS deviates from the genome average dependent on transcriptional activity of the gene, as can be seen with the averaged plots in Figure 3.2. Other regulatory factors have been shown to have meaningful nucleosome occupancy changes between conditions that correlate with a transcriptional change. It is known that occupancy around transcription factors (TFs) have a typical pattern that commonly includes nucleosome

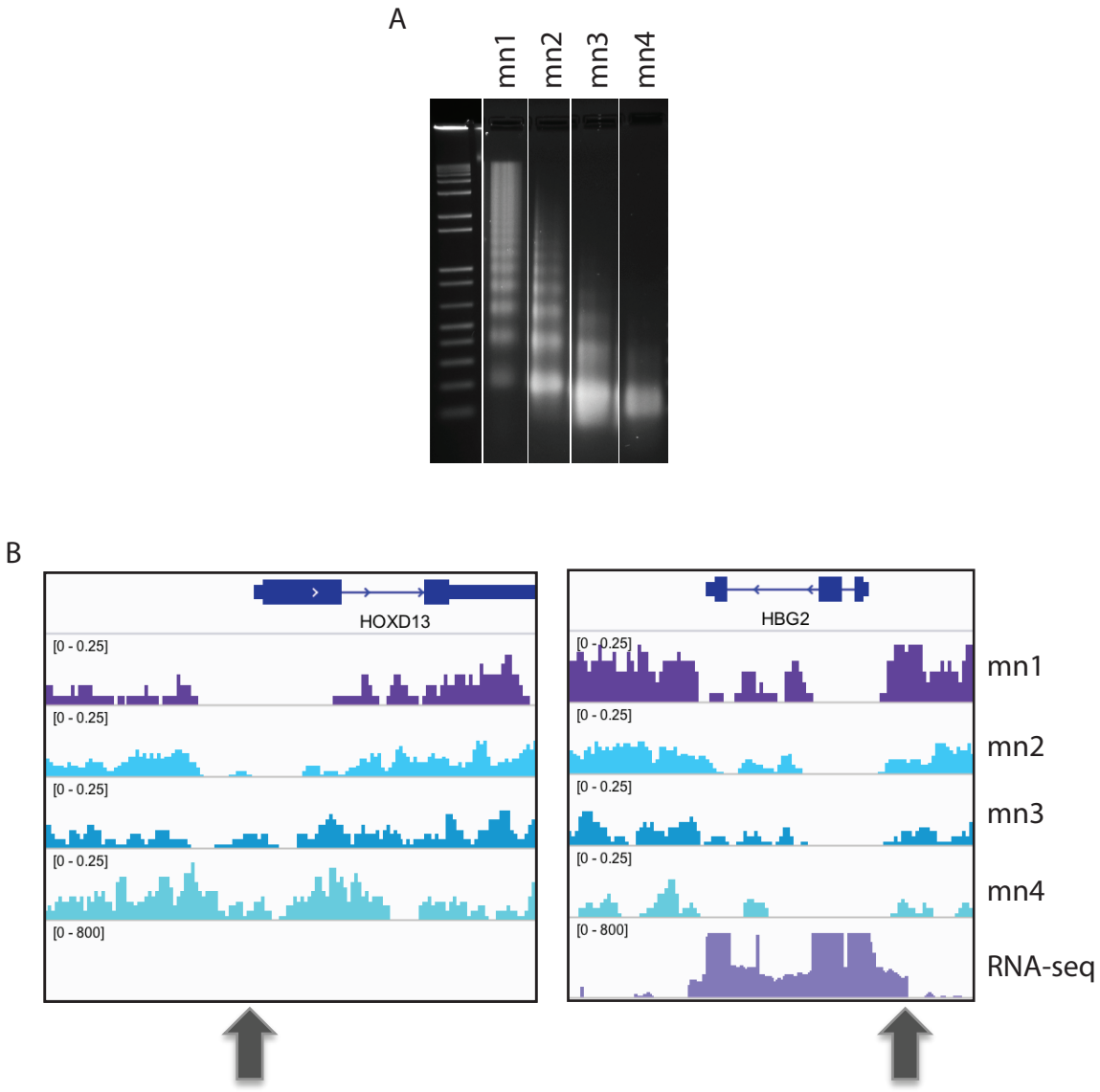



**Figure 3.3 Local patterns of nucleosome occupancy. A.** The four chosen titration ladders; mn1 is the lowest concentration of MNase used and mn4 is the highest. **B.** Nucleosome occupancy profiles obtained with four different MNase concentrations at a silenced gene, HOXD13, left, and an expressed gene, HBG2, right. Gene structure is indicated in dark blue at the top of the screenshot, and expression in the bottom track. Arrows indicate regions of interest where nucleosome occupancy increases with increasing digestion in the repressed gene promoter, and occupancy decreases with increasing expression in the active promoter.

profiles at TFs in different cell lines have been carried out using only one digestion condition[12,15,23,26]. However it can be seen in Figure 3.4A that, as with the TSS, the averaged occupancy profiles at a representative TF are highly impacted by the extent of digestion. The two digestion conditions chosen for display are well within the range of extent of digestion that has been used in publications in the field. Further, the local occupancy data shown in heatmaps beneath each averaged plot show the broad diversity of patterns that make up the averaged pattern, as well as how those patterns shift with changing MNase concentration, illustrating the heterogeneity of a collection of regions. Clearly, comparing two different samples with one digestion condition that is not perfectly matched between them obscures any ability to do local comparisons, and may possibly alter major genome-wide conclusions as well.

The genome-wide location of a plethora of DNA binding proteins has been determined and is available in public repositories, here we use data from K562 which is a tier 1 cell line in the ENCODE project[2]. An investigation of the averaged nucleosome occupancy plots of multiple types of binding proteins in K562 cells provides an interesting observation about protected DNA fragments during MNase digestion, and thus the accessibility of the DNA.

MNase titration series methodology provides a new way of looking at the physical state of a genomic region. MNase titration data can be compared to maps of the binding of proteins known to correlate with a particular transcription state or of a particular 'openness'. For example, histone 3 lysine 4 methylation (H3K4me1) is a covalent histone modification associated with enhancers, which in turn have been characterized as open[28]. Digestion at low concentrations of MNase shows a clear enrichment of protected fragments at the H3K4me1 site (center of plot, Figure 3.5A). With increasing digestion this enrichment is decreased. Any

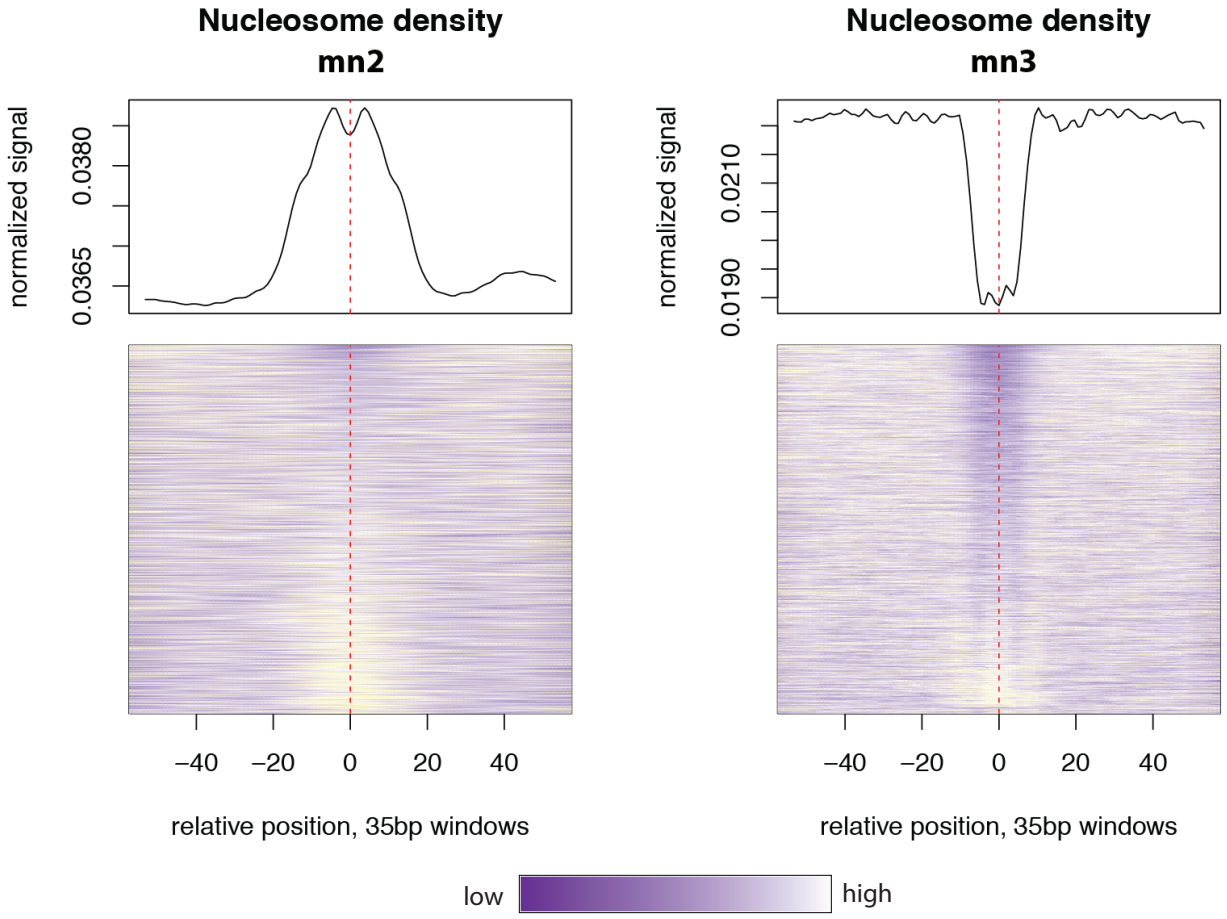



**Figure 3.4 Extent of digestion impacts global and local analysis.** Top panels: Averaged nucleosome occupancy around the Arid3 transcription factor for mn2 and mn3 samples (the middle digestions of the series of four). Bottom panels: Heatmap containing 35bp bins of nucleosome occupancy data for each Arid3 locus underlying the averaged plot above. The binding sites are centered at zero. White is high coverage and dark purple is low or no coverage.

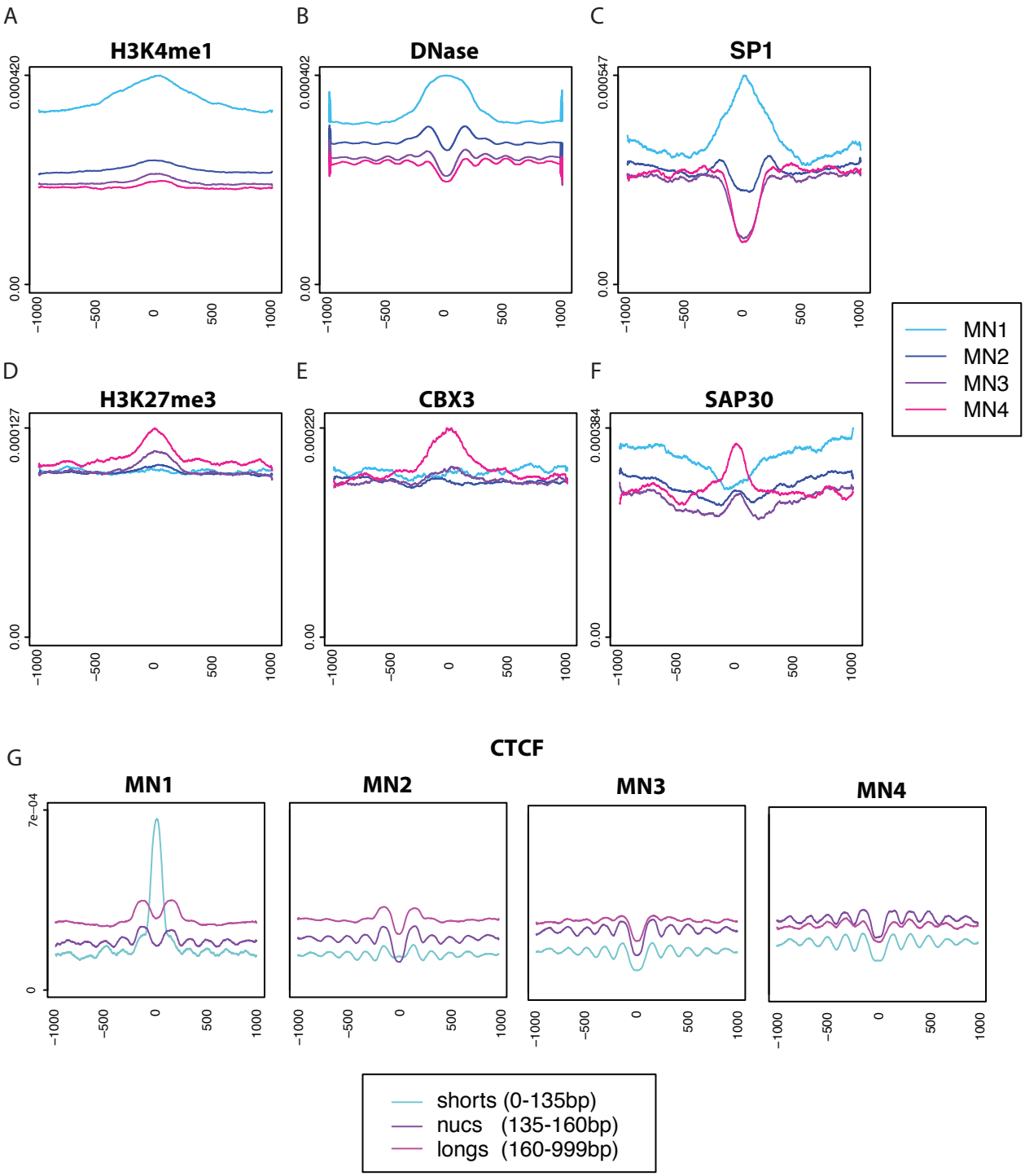



**Figure 3.5 Accessibility at regions with evidence of regulatory function A-F.** Averaged nucleosome occupancy plots around genomic loci. **A-C** are regions associated with proteins thought to have a role in open chromatin, **D-F** are regions that are associated with proteins thought to have a role in repressed chromatin. MNase sample data is colored as in the legend. **G.** Nucleosome occupancy averaged around CTCF sites, segmented by protected fragment length. Each plot is drawn to the same scale, demonstrating the strong shift of the profiles of the lowestand highest MNase conditions.

single concentration of MNase would have shown modest to moderate occupancy at these sites, however when taken together it is clear that protected fragments are liberated easily at these locations. To further analyze regions of chromatin that are considered open we assessed DNaseI sites. Regions easily digested by DNase I have been shown to largely be enhancers that are open to regulatory factor binding. MNase titration analysis shows the situation to be more complex (Figure 3.5B). In addition to phasing emanating from the DNaseI site in all titration points and decreasing occupancy with increasing digestion, low concentrations of MNase produce an averaged occupancy map with a striking enrichment at the center of DNaseI site. This seems to conflict with the premise of these sites being very open, but in fact it has also been shown that enhancers are enriched for histone variant and histone mark containing nucleosomes[9,16]. These specialized nucleosome components have fast turnover and it may be that single-concentration MNase experiments are incomplete in their mapping of nucleosome occupancy. The use of MNase titrations captures a more extensive description of a locus.

In addition to histone tail modifications and nuclease-sensitive sites, transcription factors have been shown to exert an effect on chromatin and open a closed locus. Sp1 is a transcription factor that is involved in the activation of the GATA-1 erythroid promoter in K562 cells[29]. If SP1 binding is lost chromatin accessibility at its binding site is lost, and SP1 knock down can affect the transcriptional state of the genes it binds[30]. Here, as with the other open sites discussed above (Figure 3.5A,B) the averaged nucleosome occupancy plot of SP1 shows a highly occupied region at the TF binding site under low digestion conditions, a moderately depleted region in the next highest level of MNase digestion (blue line), and severe depletion in the two highest digestion conditions (purple and pink). Again, liberating protected fragments

with the lowest concentration of MNase and then seeing a reduction with progressively higher MNase concentrations may indicate an open physical conformation.

MNase and other nucleases have been used to probe for open regions in chromatin before, however tools for assessing closed chromatin at the genome level are less common. To address the ability of the MNase titration method to describe these regions we assessed averaged nucleosome occupancy plots for marks associated with repressed or closed chromatin. It is known that H3k27me3 is deposited by Polycomb Repressive Complex 2 and, except in the case of pluripotent cells, is associated with a repressed state. The series of averaged profiles of H3K27me3 sites shows a much different pattern than do the open marks (Figures 3.5A-C), with averaged nucleosome occupancy highest in the most extensively digested titration, and lowest in the least digested sample (Figure 3.5D). This suggests that more protected fragments are liberated as the digestion progresses, which we interpret to mean the locus is refractory to digestion, or more inaccessible.

Cbx3, also known as HP1gamma, plays a role in silencing carried out by heterochromatic complexes. Cbx3 has been shown to bind H3K9me3, a repressive mark, although the exact mechanism of repression is incompletely understood[31,32]. The averaged pattern of nucleosome occupancy seen at CBX3 sites (Figure 3.5E) clearly shows no enrichment of protected DNA fragments above the surrounding regions at lower concentrations, however, a peak forms in the highest concentration. This shows that the DNA was not accessible, perhaps packed in condensed chromatin, until a sufficient amount of nuclease was added to liberate it. Lastly, as an example of occupancy at a collection of regions that may undergo deacetylation and repression through the binding of a complex containing the Sap30 protein[33] see Figure 3.5F.

The above assessment of a variety of DNA binding proteins and chromatin features, including transcription factors, histone tail marks, chromatin proteins, and remodeling complex subunits, shows that different extents of digestion have an impact on how MNase-seq data correlates with those features. It is interesting that repressive marks correlate with nucleosomes becoming apparent at high concentrations of MNase, while active marks correlate with nucleosomes mapped with low MNase concentrations.

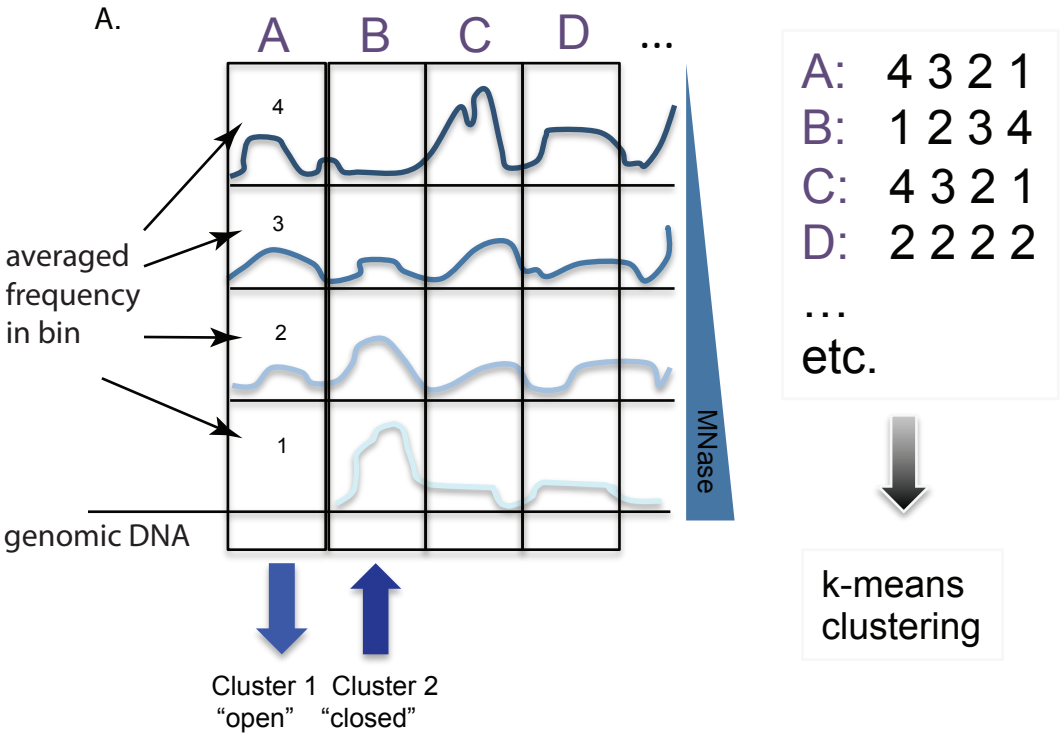
In addition to discerning open and closed chromatin structure, mapping nucleosome occupancy using MNase titrations addresses the problem of bias introduced by cutting agarose gels. It is known that nucleosome particles can have different sizes depending on the variant incorporated, for example H2AZ-containing nucleosomes are smaller than the canonical nucleosome at 120 base pairs in length[34]. Gel extraction could exclude these fragments. Additionally general experimenter error in cutting a gel could impact results. Previous work shows that fragments of different size can be useful in understanding chromatin structure. ATAC-seq, a method using transposons to probe active chromatin, demonstrated that binning of fragment size sheds light on transcription factor binding[35]. Additionally, analysis of a range of fragment sizes liberated from an MNase-seq experiment following salt fractionation suggests small fragments map to known TF binding sites[36]. Using SPRI bead size selection and retaining fragments sizes outside of the 150 base pair canonical nucleosome size allows us to answer questions about the size distribution patterns at regions of interest.

It is interesting that many published nucleosome occupancy maps show nucleosome-depleted regions (NDRs) at protein binding sites. For example, CTCF has a striking pattern of nucleosome occupancy averaged around its locations in the genome. It is characterized by a

NDR flanked by at least 5 well-positioned nucleosomes on either side[37]. Here, when the CTCF averaged profiles are created after binning by length into short, nucleosome-length and long groups, an interesting pattern emerges, see Figure 3.5G. The characteristic pattern of depleted CTCF binding site flanked by phased nucleosomes is seen in most levels of digestion and with most fragment lengths. Notably, the small fragments in the lightest digestion condition show a strikingly different phenotype: a sharp peak at the binding site. This peak could mean several things, it could be caused by protection of a smaller than average nucleosome or, most likely, of CTCF or another non-histone protein binding. The methodology described here allows us to probe differently accessible chromatin, whether active or not, across the genome.

## Two classes of chromatin

MNase digestion series produce interesting patterns of occupancy averaged at known chromatin features. Generally two major categories of chromatin are discussed--open and closed. It appears from the alignments in Figure 3 that we may be able to harness the patterns in a digestion series to categorize the physical state of chromatin not just at known features but across the genome. To categorize genomic regions according to the pattern seen in the titration series we computed frequencies of the digestion fragments in 300-bp non-overlapping bins genome-wide for each titration point (resulting in four numbers per bin) and applied k-means clustering to these frequencies. An schematic example of what one locus with several bins looks like in this analysis is shown in Figure 3.6A. The k-means clustering is an unsupervised method to partition data into a specified number of clusters, k, and using k=2 resulted in the clusters with opposite patterns of mean frequency values. We interpret these clusters as potentially



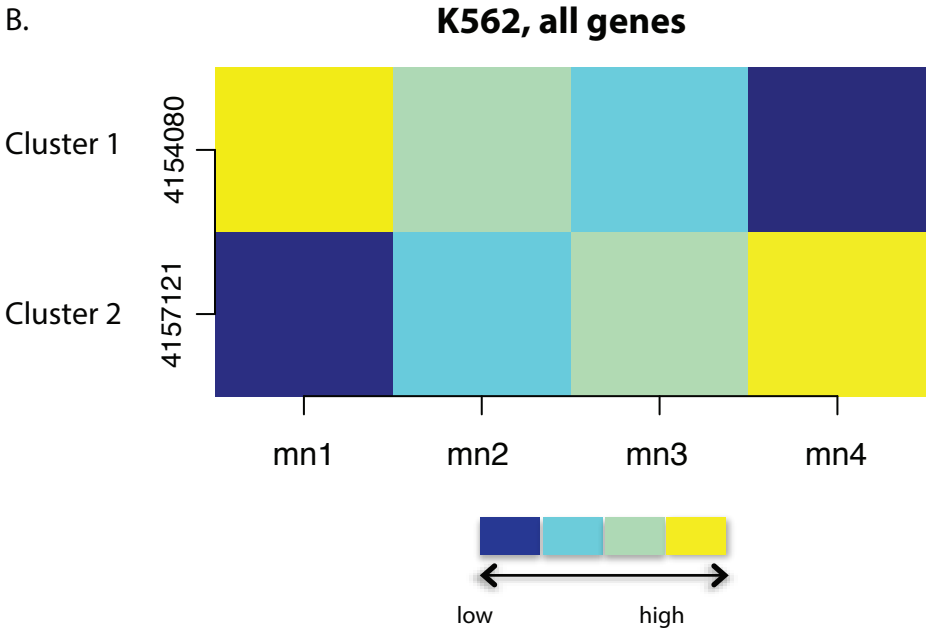




**Figure 3.6 Genome-wide k-means clustering. A.** Genomic data is segmented into bins (A, B, C, D) for each of the four MNase conditions. Average nucleosome occupancy per bin for each locus is treated as a vector in a k-means clustering algorithm. **B.** Heatmap of resulting k-means clusters. Rows correspond to clusters and columns to individual MNase concentrations. Blue and yellow correspond to low and high signal, respectively. Numbers of bins in each cluster are on the left side of the heatmap. Cluster 1 includes regions that decrease in nucleosome occupancy with increasing digestion, and Cluster 2 has the opposite pattern with increasing digestion.

corresponding to open and closed chromatin and test a few predictions of that hypothesis below.

Our results reveal that we can distinguish two major scenarios of the response of chromatin to MNase titration series, which likely correspond to accessible and inaccessible loci. To see how these clusters of regions compare with known marks of chromatin state we determined their relationship with data available from the ENCODE project. Cluster regions were assessed for overlap with twelve commonly studied histone marks, transcription factors and remodeling and other chromatin proteins (Figure 3.7A). The fraction of each cluster overlapping a mark is depicted in a bar chart for each protein shown in Figure 3.7B. Most marks show a stronger association with the open cluster (cluster 1), with H3K27ac H3K4me3 and H3K9ac showing the strongest effect. H3K9me3, a known repressive mark shows enrichment in the closed cluster, cluster 2. Interestingly, H3K36me3, a mark seen in actively transcribed gene bodies, shows enrichment in closed regions. H3K36me3 plays a role in preventing histone exchange after the passage of RNA polymerase II so that the proper reassembly of context-specific nucleosomes can occur. It does this by association with remodelers that space nucleosomes and by recruiting deacetylases[38]. It follows that this mark may exist in a locally closed chromatin conformation, despite being localized in actively transcribed regions. This finding shows the importance of a metric that assesses the physical state of chromatin when studying the mechanisms of effects of chromatin proteins and modifications.

Clusters show correlation with marks previously associated with particular chromatin states, and this data taken together deepens our understanding of regulation in these regions. However, the above analysis is limited to regions already well-studied. To determine what

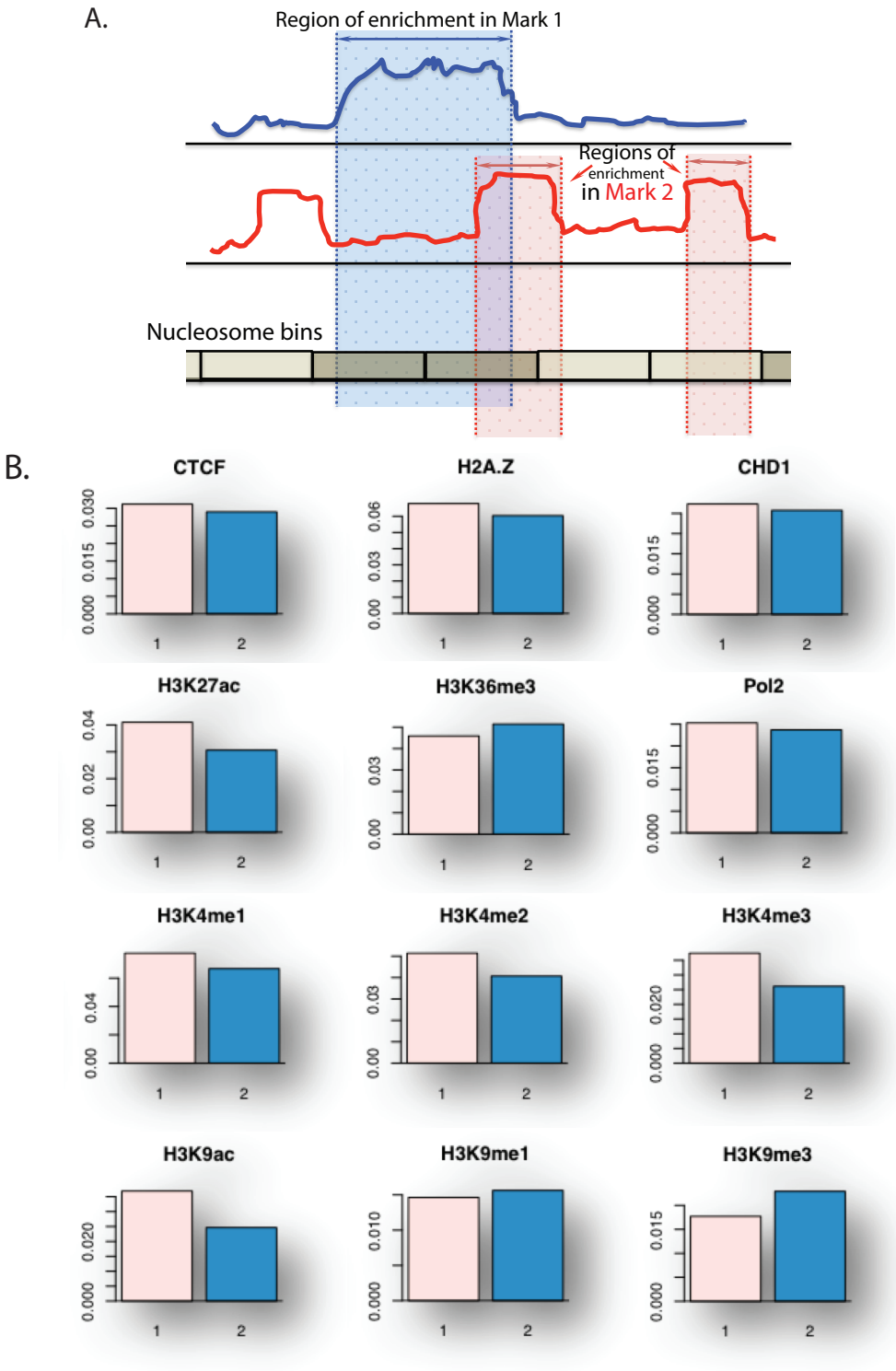



**Figure 3.7 Relationship between nucleosome occupancy and mapped chromatin features. A.** Schematic showing how relationship between occupancy and locations of marks profiled in K562 as part of the ENCODE project **B.** Percentage of overlap of Cluster 1 (pink, open) and cluster 2 (blue, closed) bins with chromatin proteins and covalent histone tail modifications.

MNase titration data tells us across the genome, we developed a quantitative measure of the accessibility for any given locus, described below.

## Quantitative profiling of genome-wide chromatin accessibility

To generate a quantitative measure of the response of any particular genomic locus to MNase digestion, we introduce a metric that reflects the rate of the change in digestion fragment counts at a given locus upon titrating MNase concentrations (see a representative hypothetical locus in Figure 3.8A). Individual loci in the genome respond differently to MNase digestion depending on the nucleosome binding at that location, as demonstrated in yeast in Bryant et al.[22], and to our knowledge this is the first quantitative assessment of the response of nucleosomal DNA to a range of MNase concentrations in mammals, or across a genome. This metric, which we termed 'MNase accessibility' or MACC, represents the slope of the linear regression fitted to the digestion frequencies obtained for each MNase concentration (logarithmic scale of MNase concentrations was used to obtain equidistant distribution of experimental points). Essentially, the slope of the occupancy scores will tell us if occupancy coverage at a locus is increasing or decreasing with changing digestion. We use this pattern to characterize loci as open or closed.

The MACC score appears to be correlated with the GC-content of the underlying DNA sequence in both human and fly data (Figure 3.8B, *D. melanogaster* data shown). Such correlation can be partly explained by the known dependence of the nucleosome stability and observed occupancy on the GC-content of DNA wrapped around histone core[39,40]. This association with GC content is expected since the nucleosome prefers certain sequences to others, and whether nucleosome sit in their preferred locations or have been remodeled to

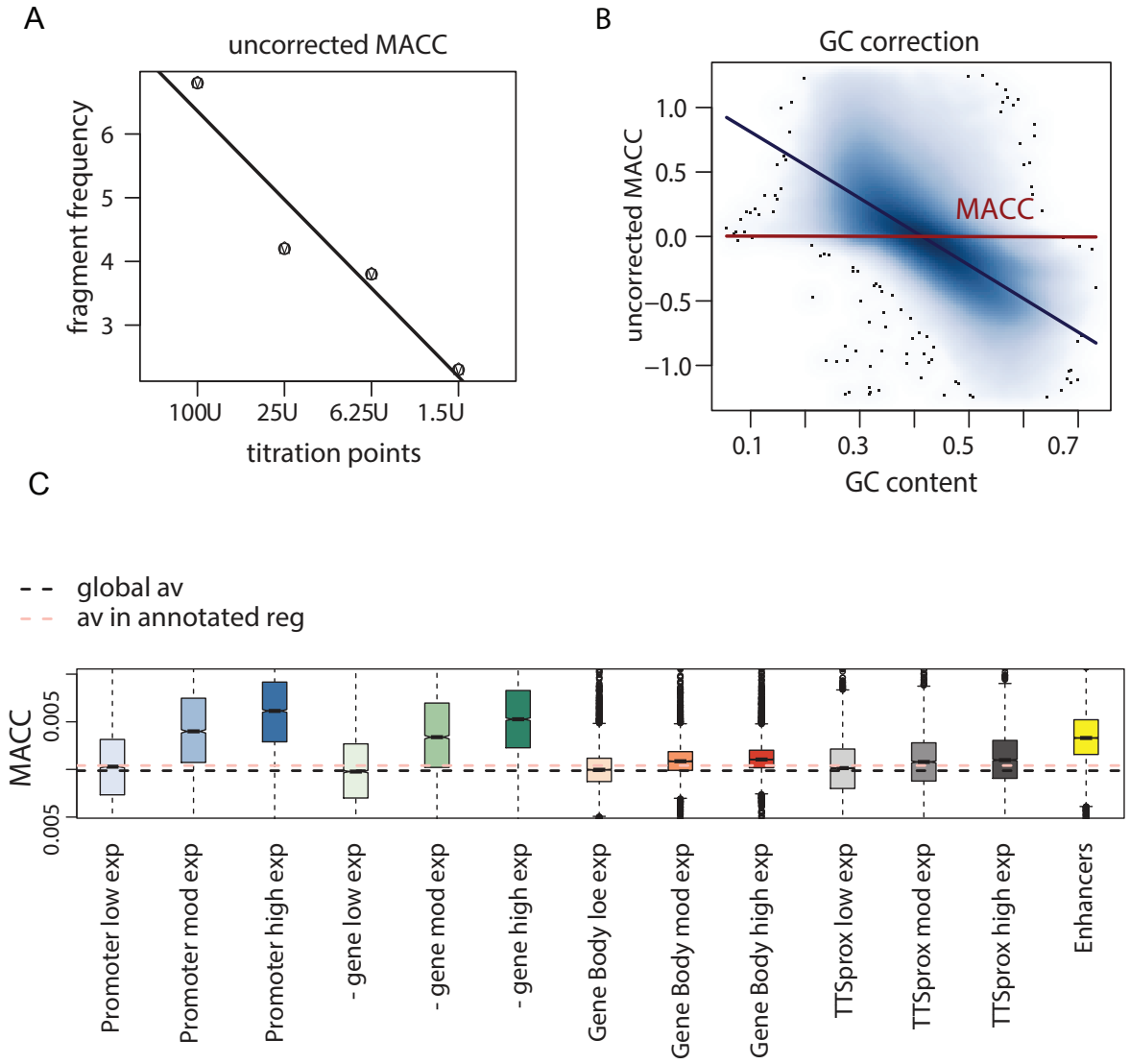



**Figure 3.8 Computation and of MNase accessibility (MACC). A.** Fit of the the linear regression model to counts of digestion fragments produced at four MNase concentrations aligned to a specific 300-bp bin. **B.** Correction for the GC-bias in the slopes computed with the linear regressions. The lowess-based correction was preformed on genome-wide distribution of the MACC values (blue) and resulted in removal the dependence of these values on the GC-content of underlying DNA sequence. The black and red lines represent correlation trends in uncorrected and corrected data, respectively. **C**. The distribution of GC-corrected MACC values within annotated regions. Results shown for promoters (1kb upstream of TSS, blue), 5'-ends of genes (1kb downstream of TSS, green), gene bodies (red), regions around transcription termination sites (+/-1kb, grey), and enhancers. The shade of the color within each group of regions represents the magnitude of expression level. Black dashed line presents genome-wide mean for MACC value, and pink dashed line represents the mean for MACC value within annotated regions only. For the overlapping regions the following priority rule was used: enhancer > promoters > 5'-gene > TTS-prox > gene bodies.

other locations is of interest. However, MNase also has a known cutting bias[19]. Because the degree of the dependence of nucleosome stability and accessibility on GC-content is hard to estimate due to the complexity of the regulatory forces on the epigenome, we chose to completely remove this dependence using loess-based normalization to avoid possible artifacts in both human and fly data. The GC content of bins across the genome was computed and used to adjust the MACC score in each bin such that the resulting correlation plot of GC content and MACC scores across the genome has a slope of zero (see Methods for detail). The effect of such normalization is illustrated in Figure 3.8B. We note that due to the fact that nucleosome positioning is somewhat impacted by sequence content the above manipulation may lead to over-correction of the data, which would in turn result in conservative estimate of the MACC scores. In addition, human and fly have very different GpC distribution and use, and it is of interest whether this correction has a similar impact on findings in both organisms.

We investigated how the MACC scores are distributed in the regulatory regions of the genome (Figure 3.8C). High MACC scores (positive slope) indicate low nucleosome occupancy at high concentrations of MNase, and high nucleosome occupancy at low concentrations of MNase. Low MACC scores indicate high occupancy at high concentrations, and low occupancy at low concentrations. The highest positive mean values of MACC can be detected for the regions associated with active transcription and specifically for gene promoters and active enhancers. Gene bodies also show higher accessibility when transcribed, but this dependence is less pronounced than it is in promoters. The 'canonical' nucleosome-depleted region is located slightly upstream of TSS, and given the differences in occupancy profiles across MNase digestions at this location (Figure 3.2) we asked if accessibility in both upstream and

downstream TSS-proximal regions (referred here as “promoters” and “5’-regions of genes”) is important for active transcription. This analysis reveals that presence of an ‘accessible state’ in at least one of these regions is associated with increase in gene expression, however, presence of ‘accessible state’ in both regions leads to a significantly larger gain in the transcriptional output (Figure 3.9A). The level of DNA accessibility in entire TSS-proximal regions quantitatively correlates with gene expression. We also observed that accessibility of the promoters tends to have higher values than that of 5’-regions of the same genes, however this trend is associated with the levels of gene expression only modestly (Figure 3.9B).

## Nucleosome accessibility and occupancy on a broad and fine scale

Our results demonstrate that the MACC metric provides useful information both on a larger scale of domain structure of chromatin, allowing broad detection of changes in DNA accessibility associated with gene activation or silencing, and on a smaller scale allowing detection of regions impacted by binding of regulatory proteins. We observe that regions carrying distinct sets of histone marks differ in their physical chromatin properties. Taken together these results provide an insight into molecular mechanisms connecting placement of histone modifications and gene regulation. An integrated example of lessons learned using MNase titration methodology at a HOX locus can be seen in Figure 3.10. The HOX gene clusters have been characterized as being epigenetically repressed during development by way of physical compaction or organization by Polycomb group proteins[41]. Both k means-produced clusters and MACC scores plotted along the region show broad scale physical domains, with open regions correlating with transcription and classically open histone tail modification marks

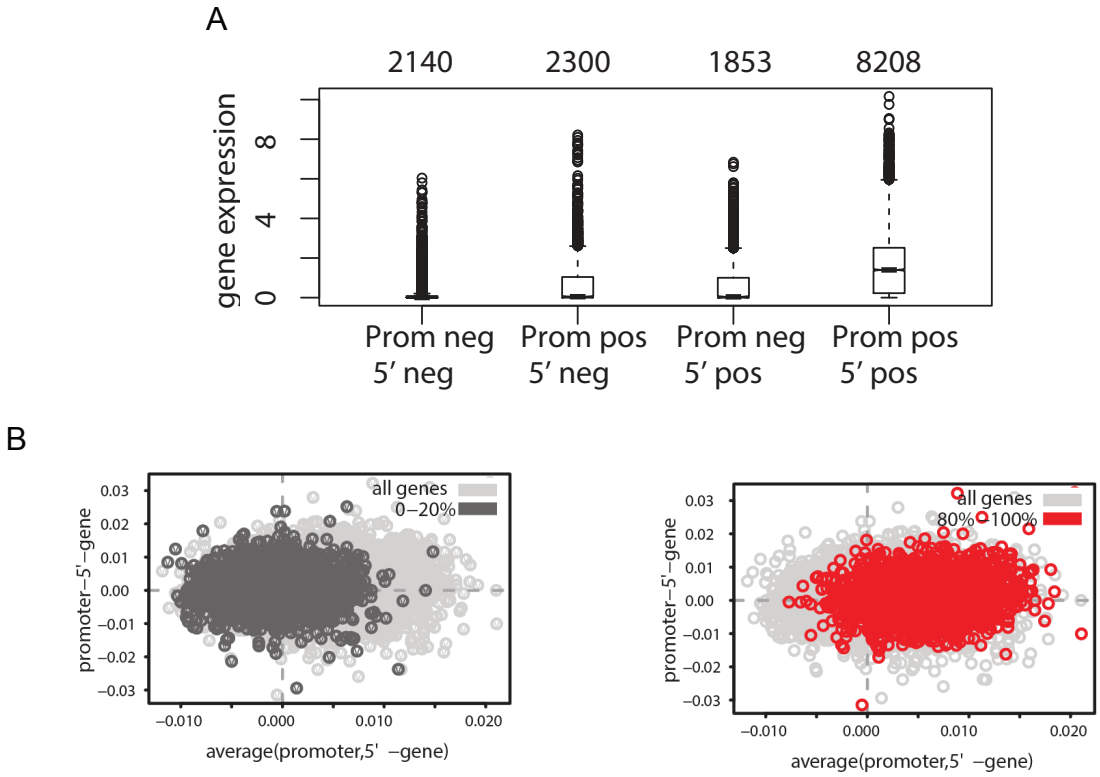



**Figure 3.9 Role of DNA accessibility in transcriptional regulation. A.** Each gene was characterized by two accessibility scores, corresponding to its promoter and 5′-end of the gene genes (regions 1kb up- and down-stream of TSS respectively). The genes were placed in four groups according to the possible combinations of the signs of these scores (positive "accessible" or negative "inaccessible"), and the distributions of the expression values in each of these groups are shown in the plot. **B.** Relation between the accessibility of promoters and 5′-ends of the individual genes. The plots show data for all genes (light grey) and for the two non-overlapping groups of genes stratified by the expression level (dark gray corresponds to the genes with lowest expression level, 0-20%, and red corresponds to genes with 80-100% of the expression range).

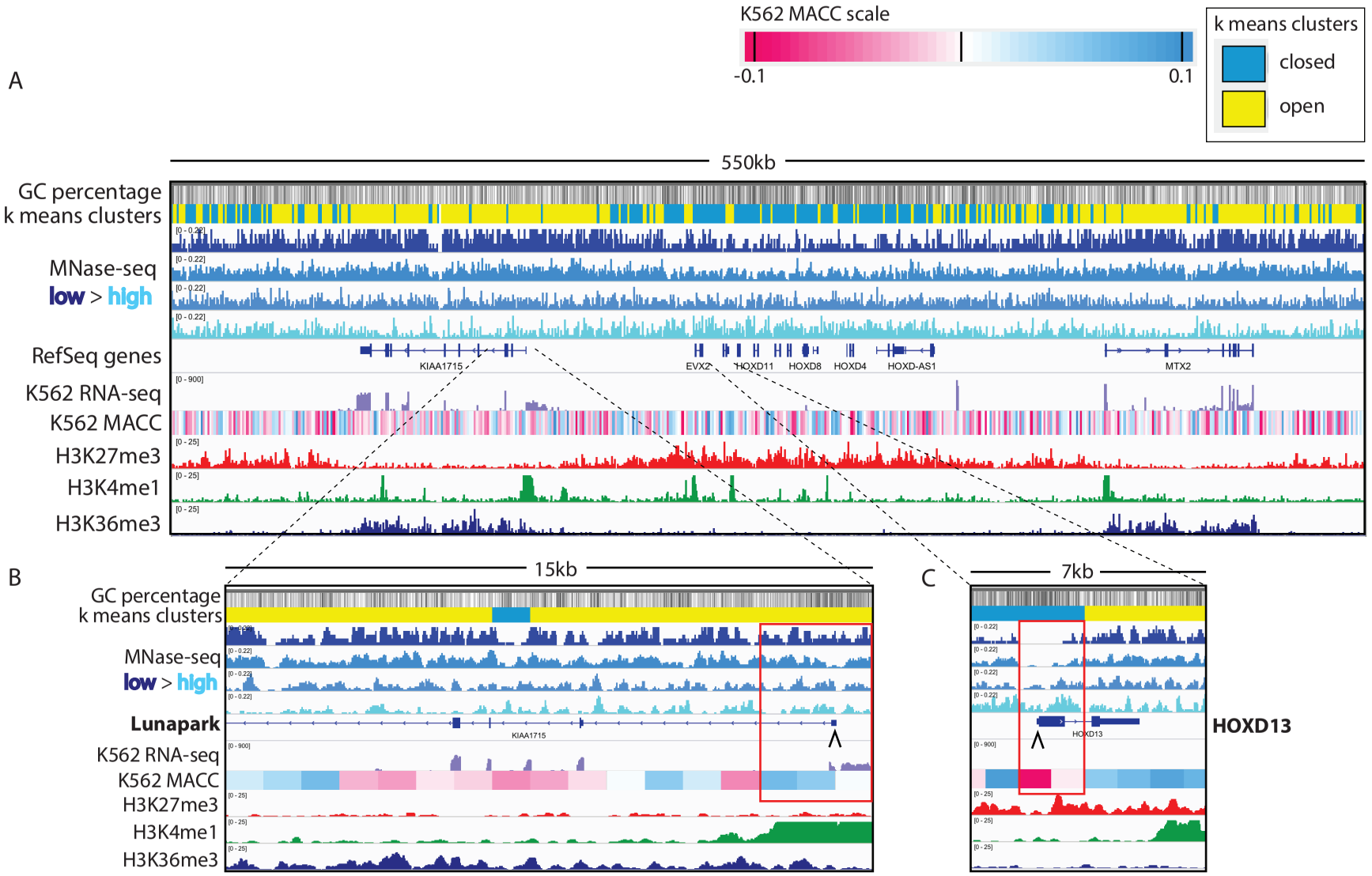



**Figure 3.10 Global and local representation of kmeans clusters and MACC scores** A. A 550kb region surrounding the HOXD locus. Scale for kmeans cluster and MACC scores is shown, top right. **B.** A transcribed gene, Lunapark, that sits outside the HOX locus. TSS is 'open'., red box. **C.** A repressed gene, HOXD13. TSS is 'closed', red box.

and closed regions showing no transcription and repressive marks. An actively transcribed locus and a repressed locus are shown with greater detail in Figure 3.10B and C, respectively.

Taken together these results indicate that our approach can provide both quantitative and qualitative description of chromatin accessibility, demonstrating that both the 'active' and 'repressed' regions of the genome show physical characteristics of being open and closed. MNase titration followed by whole genome sequencing is novel in its ability to profile active and repressed chromatin accessibility on a local and genome-scale level.

## Discussion

The location of nucleosomes in chromatin impacts the accessibility of DNA to regulatory factors, and thus mapping the nucleosome occupancy of a genome should tell us about its regulatory state. The first genome-wide nucleosome occupancy maps produced using MNase-seq in mammals provided a description of the global characteristics of regulatory regions, including transcription start sites, enhancers and transcription factor binding sites. However, nucleosome occupancy maps created with one condition are not easy to interpret as they reflect only a fraction of the nucleosomes in chromatin. Further, even well-matched digestion conditions may not be appropriate when dealing with a comparison of cell states that have fundamentally different chromatin architecture. For example, pluripotent cells have open chromatin, and if this chromatin is digested with the same conditions as a differentiated cell population's chromatin is digested, the results will be very different subpopulations of chromatin.

MNase titration assays address these issues by profiling both DNA accessibility and nucleosome occupancy. MNase liberates different populations of protected DNA fragments

with different levels of digestion. Low digestion levels produce occupancy maps with enrichment at regions typically shown to be DNaseI-sensitive, thought to be of regulatory function, and to be nucleosome-free. Samples digested at high levels no longer contain many of the fragments liberated in the low digest, but now show enriched occupancy at regions that weren't occupied in the low nuclease conditions. Further, when nucleosome occupancy in these different conditions is averaged around known active and repressed regions these patterns of accessible and inaccessible fragments correlate with the activity of the regions. We take these digestion patterns and correlation with known marks of active and repressed regions to mean that low digestion profiles very accessible sites, and high digestion profiles inaccessible sites, and that any single digestion condition misses part of this picture.

The new metric reported here, MACC, quantifies accessibility in addition to nucleosome occupancy. MACC compares well with mapped histone marks and other metrics describing physical properties of chromatin, and also provides a metric to compare as-of-yet uncharacterized regions of the genome. MNase titration is easy to perform, with just one enzymatic assay to carry out with varying amounts of enzyme. This is contrasted to nucleosome turnover experiments or FRAP which require time-consuming genomic modifications, labeling reactions that are often incomplete or biased, or microscopy which is also open to experimenter bias and is laborious. Additionally, this assay can be performed on chromatin from any cell type, and is independent from antibody availability. MACC works in the genomes of different GC-composition and complexity, as all analysis performed on human K562 chromatin was also performed in *D. melanogaster*, and the major conclusions were the same.

Thus, MACC is novel metric, which produces useful chromatin accessibility information on local and regional scales and in multiple chromatin states.

## Methods

### Experimental Procedures

Cell Culture

K562 cells were grown in IMDM with 10% FBS, pen/strep, and 2mM glutamine, at a density between $1x10^5$ and $1x10^6$ cells, at 5% $CO_2$ and at 37 degrees C. S2 cells were grown in Schneider's medium plus 10% FBS and pen/strep, at 5% $CO_2$ and at 37 degrees C.

Digestion with MNase

Human cells were expanded to yield approximately 4 million cells/reaction and cross-linked with 1.1% formaldehyde for 10 minutes at room temperature. After quenching nuclei were isolated and treated with a range of 18 MNase concentrations for 15 minutes at room temperature. EDTA and EGTA were added to stop the digestion. Cross-link reversal was performed at 65$^{o}$C for at least 16 hours followed by an RNase for 30 minutes at 37 degrees C and subsequent proteinase K digestion overnight, 55 degrees C. DNA was purified by phenol-chloroform extraction. Ampure SPRI beads (Beckman Coulter) were used in a double size selection with ratios of 0.7X and 1.7X to obtain a range of fragment sizes from approximately 100 bp to 1000 bp. The resulting fragments from four MNase concentrations in the range were prepared individually for barcoded sequencing on an Illumina HiSeq instrument.
*D. melanogaster* S2 cells were used at 1 million cells/reaction, were crosslinked with 1.1% formaldehyde for 10 minutes at room temperature and were treated with a range of four

MNase concentrations at 37 degrees C for 3 minutes. DNA clean-up was achieved with RNase at 37C for 30 min, and subsequent proteinase K digestion at 55C for 1 hr, and cross link reversal at 65C for 1 hour. DNA was purified by phenol-chloroform extraction. Ampure SPRI beads (Beckman Coulter) were used in a double size selection with ratios of 0.6X and 1.8X to obtain a range of fragment sizes from approximately 100 bp to 1000 bp.

Illumina HiSeq Library preparation and sequencing

100 ng of mononucleosome DNA (human), or approximately 1ng DNA (fly), was used for library preparation, with limited numbers of PCR amplification rounds[42], and genomic alignments of paired-end 50 bp reads were performed using Bowtie[43] followed by further tag processing and filtering with the SPP workflow[44]. All alignments and annotations used the human genome assembly hg19 and the *D melanogaster* genome assembly dm3.

## Bioinformatic and statistical data analysis

Sequencing data preprocessing and initial analysis

Sequenced 50-bp paired-end tags were mapped to the human genome (hg19) or fly genome dm3 for the corresponding cell types using the Bowtie aligner v. 0.12.7[43]. Only uniquely mapped tags with no more than two mismatches in the first 28 bp of the tag were retained. Genomic positions with the numbers of mapped tags above the significance threshold of z-score=7 were identified as anomalous, and the tags mapped to such positions were discarded. The coordinates of the genes were taken according to the annotations for hg19 and dm3 versions of the fly human genomes respectively. Normalization was performed to remove

sequencing coverage differences between samples by calculating frequency as reads per million mapped. The gene proximal profiles were calculated and plotted as described previously[45,46]

Averaged nucleosome occupancy plots

Frequency values in regions flanking the midpoint of mapped DNA binding protein peaks were averaged on a base pair scale. Mapped DNA binding protein peaks were derived from data downloaded from UCSC, http://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/. The average values were plotted with the centerpoint at 0 on the X axis and average frequency on the Y axis using the R program (http://www.r-project.org).

Clustering

Genome-wide nucleosome occupancy frequency data was partitioned into 300 base pair (fly) and 1000 base pair (human) non-overlapping bins and averaged. This was performed separately for each MNase titration point. Every genomic bin is then associated with four values, and these four numbers were used in k-means clustering (R, stats package) with k=2 to these frequencies.

Gene expression

RNA seq data for K562 cells from the ENCODE[2] project was used. Expression level was determined by plotting the log expression (RPKM) against the density of coverage for each gene. The minimum density count in the resulting bimodal plot determined the threshold of high expression and low expression of genes.

MACC

Genome-wide nucleosome occupancy frequency data was partitioned into 300 base pair (fly) and 1000 base pair (human) non-overlapping bins and averaged. Each MNase concentration has a value in each genomic bin. The logarithmic scale of MNase concentrations was used to obtain equidistant distribution of experimental points. These experimental points and the average digestion frequencies obtained for each MNase concentration considered together to compute a slope, which we term a 'MACC score'.

GC correction

The GC content in each genomic bin (1kb for human, 300bp for fly, as above) was calculated, and a Loess curve was computed for this GC content and MACC scores for all bins. GC-normalized MACC scores were computed by subtracting the loess trend value for the corresponding GC content from the MACC score for each bin.

## Acknowledgments

## Contributions

A.C. performed human cell line experiments, S.B. performed drosophila melanogaster cell line experiments, M.Y.T. analyzed the data, A.C., S.B., M.Y.T., and R.E.K. designed the study, interpreted the results and wrote the paper. All authors read and contributed editing to the manuscript during its preparation.

## References


1 Almer, A. & Hörz, W. Nuclease hypersensitive regions with adjacent positioned nucleosomes mark the gene boundaries of the PHO5/PHO3 locus in yeast. *The EMBO journal* **5**, 2681-2687 (1986).

2 Consortium, E. P. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS biology* **9**, doi:10.1371/journal.pbio.1001046 (2011).

3 Wang, Q. *et al.* CR Cistrome: a ChIP-Seq database for chromatin regulators and histone modification linkages in human and mouse. *Nucleic acids research* **42**, 8, doi:10.1093/nar/gkt1151 (2013).

4 Narlikar, G. J., Fan, H.-Y. Y. & Kingston, R. E. Cooperation between complexes that regulate chromatin structure and transcription. *Cell* **108**, 475-487 (2002).

5 Armache, K.-J. J., Garlick, J. D., Canzio, D., Narlikar, G. J. & Kingston, R. E. Structural basis of silencing: Sir3 BAH domain in complex with a nucleosome at 3.0 Å resolution. *Science (New York, N.Y.)* **334**, 977-982, doi:10.1126/science.1210915 (2011).

6 Meshorer, E. *et al.* Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Developmental cell* **10**, 105-116, doi:10.1016/j.devcel.2005.10.017 (2006).

7 Misteli, T., Gunjan, A., Hock, R., Bustin, M. & Brown, D. T. Dynamic binding of histone H1 to chromatin in living cells. *Nature* **408**, 877-881, doi:10.1038/35048610 (2000).

8 Skene, P. & Henikoff, S. Histone variants in pluripotency and disease. *Development (Cambridge, England)* **140**, 2513-2524, doi:10.1242/dev.091439 (2013).

9 Jin, C. *et al.* H3.3/H2A.Z double variant-containing nucleosomes mark 'nucleosome-free regions' of active promoters and other regulatory regions. *Nature genetics* **41**, 941-945, doi:10.1038/ng.409 (2009).

10 Widom, J. Role of DNA sequence in nucleosome stability and dynamics. *Q Rev Biophys* **34**, 269-324 (2001).

11 Tolstorukov, M., Colasanti, A., McCandlish, D., Olson, W. & Zhurkin, V. A novel roll-and-slide mechanism of DNA folding in chromatin: implications for nucleosome positioning. *Journal of molecular biology* **371**, 725-738, doi:10.1016/j.jmb.2007.05.048 (2007).

12 Teif, V. *et al.* Genome-wide nucleosome positioning during embryonic stem cell development. *Nature structural & molecular biology* **19**, 1185-1192, doi:10.1038/nsmb.2419 (2012).

13 Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132**, 887-898 (2008).

14 Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75-82, doi:10.1038/nature11232 (2012).

15 Gaffney, D. *et al.* Controls of nucleosome positioning in the human genome. *PLoS genetics* **8**, doi:10.1371/journal.pgen.1003036 (2012).

16 Henikoff, S., Henikoff, J. G., Sakai, A., Loeb, G. B. & Ahmad, K. Genome-wide profiling of salt fractions maps physical properties of chromatin. *Genome research* **19**, 460-469, doi:10.1101/gr.087619.108 (2009).

17 Chung, H.-R. *et al.* The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS One* **5**, doi:10.1371/journal.pone.0015754 (2010).

18 Dingwall, C., Lomonossoff, G. & Laskey, R. High sequence specificity of micrococcal nuclease. *Nucleic acids research* **9**, 2659-2673, doi:10.1093/nar/9.12.2659 (1981).

19 Hörz, W. & Altenburger, W. Sequence specific cleavage of DNA by micrococcal nuclease. *Nucleic acids research* **9**, 2643-2658, doi:10.1093/nar/9.12.2643 (1981).

20 Xi, Y., Yao, J., Chen, R., Li, W. & He, X. Nucleosome fragility reveals novel functional states of chromatin and poises genes for activation. *Genome research* **21**, 718-724, doi:10.1101/gr.117101.110 (2011).

21 Weiner, A., Hughes, A., Yassour, M., Rando, O. & Friedman, N. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome research* **20**, 90-100, doi:10.1101/gr.098509.109 (2010).

22 Bryant, G. O. *et al.* Activator control of nucleosome occupancy in activation and repression of transcription. *PLoS biology* **6**, 2928-2939, doi:10.1371/journal.pbio.0060317 (2008).

23 Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516-520, doi:10.1038/nature10002 (2011).

24 Rizzo, J., Bard, J. & Buck, M. Standardized collection of MNase-seq experiments enables unbiased dataset comparisons. *BMC molecular biology* **13**, 15, doi:10.1186/1471-2199-13-15 (2012).

25 Li, Z., Schug, J., Tuteja, G., White, P. & Kaestner, K. The nucleosome map of the mammalian liver. *Nature structural & molecular biology* **18**, 742-746, doi:10.1038/nsmb.2060 (2011).

26 Kundaje, A. *et al.* Ubiquitous heterogeneity and asymmetry of the chromatin environment at regulatory elements. *Genome research* **22**, 1735-1747, doi:10.1101/gr.136366.111 (2012).

27 Tolstorukov, M. *et al.* Swi/Snf chromatin remodeling/tumor suppressor complex establishes nucleosome occupancy at target promoters. *Proceedings of the National Academy of Sciences of the United States of America* **110**, 10165-10170, doi:10.1073/pnas.1302209110 (2013).

28 Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nature reviews. Genetics*, doi:10.1038/nrg3682 (2014).

29 Hou, C. H. *et al.* Involvement of Sp1/Sp3 in the activation of the GATA-1 erythroid promoter in K562 cells. *Cell research* **18**, 302-310, doi:10.1038/cr.2008.10 (2008).

30 Wang, J., Liu, X., Ni, P., Gu, Z. & Fan, Q. SP1 is required for basal activation and chromatin accessibility of CD151 promoter in liver cancer cells. *Biochemical and biophysical research communications* **393**, 291-296, doi:10.1016/j.bbrc.2010.01.127 (2010).

31 Ruan, J. *et al.* Structural basis of the chromodomain of Cbx3 bound to methylated peptides from histone h1 and G9a. *PloS one* **7**, doi:10.1371/journal.pone.0035376 (2011).

32 Maksakova, I. A. *et al.* H3K9me3-binding proteins are dispensable for SETDB1/H3K9me3-dependent retroviral silencing. *Epigenetics & chromatin* **4**, 12, doi:10.1186/1756-8935-4-12 (2010).

33 Zhang, Y. *et al.* SAP30, a novel protein conserved between human and yeast, is a component of a histone deacetylase complex. *Molecular cell* **1**, 1021-1031 (1998).

34 Tolstorukov, M., Kharchenko, P., Goldman, J., Kingston, R. & Park, P. Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. *Genome research* **19**, 967-977, doi:10.1101/gr.084830.108 (2009).

35 Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213-1218, doi:10.1038/nmeth.2688 (2013).

36 Teves, S. S. & Henikoff, S. Heat shock reduces stalled RNA polymerase II and nucleosome turnover genome-wide. *Genes & development* **25**, 2387-2397, doi:10.1101/gad.177675.111 (2011).

37 Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS genetics* **4**, doi:10.1371/journal.pgen.1000138 (2007).

38 Venkatesh, S. & Workman, J. L. Recognizing methylated histone variant H3.3 to prevent tumors. *Cell research* **24**, 649-650, doi:10.1038/cr.2014.50 (2014).

39 Lowary, P. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of molecular biology* **276**, 19-42, doi:10.1006/jmbi.1997.1494 (1998).

40 Dennis Jh, F. H. Y. R. S. M. Y. G. M. J. C. R. D. J. P. D. G. R. O. J. N. W. Independent and complementary methods for large-scale structural analysis of mammalian chromatin. *Genome Res* **17**, 928-939 (2007).

41 Delest, A., Sexton, T. & Cavalli, G. Polycomb: a paradigm for genome organization from one to three dimensions. *Current opinion in cell biology* **24**, 405-414, doi:10.1016/j.ceb.2012.01.008 (2012).

42 Bowman, S. K. *et al.* Multiplexed Illumina sequencing libraries from picogram quantities of DNA. *BMC Genomics* **14**, 466, doi:1471-2164-14-466 [pii] 10.1186/1471-2164-14-466 (2013).

43 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10**, R25, doi:gb-2009-10-3-r25 [pii] 10.1186/gb-2009-10-3-r25 (2009).

44 Kharchenko, P. V., Woo, C. J., Tolstorukov, M. Y., Kingston, R. E. & Park, P. J. Nucleosome positioning in human HOX gene clusters. *Genome Res* **18**, 1554-1561 (2008).

45 Tolstorukov, M. Y., Kharchenko, P. V., Goldman, J. A., Kingston, R. E. & Park, P. J. Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. *Genome research* **19**, 967-977, doi:gr.084830.108 [pii] 10.1101/gr.084830.108 (2009).

46 Tolstorukov, M. Y. *et al.* Histone variant H2A.Bbd is associated with active transcription and mRNA processing in human cells. *Molecular cell* **47**, 596-607, doi:10.1016/j.molcel.2012.06.011 (2012).

# Chapter 4 Discussion and future directions

Chromatin in the nucleus is the template upon which transcription, DNA repair and DNA replication are carried out. Many nuclear processes alter the state of chromatin, these alterations have transcriptional and cell fate consequences. The studies in this dissertation harness a powerful enzymatic tool, micrococcal nuclease (MNase), to profile the physical state of chromatin across the genomes of several mammalian cell types.

Historically MNase has been used to digest linker DNA between nucleosomes leaving a population of protected fragments for further analysis. We used MNase to probe the chromatin of two cell types known to have very different chromatin features: pluripotent and differentiated cells. We expected large global differences between the two cell types given the 'open' chromatin state of pluripotent cells and the 'closing' of loci due to differentiation. Our maps of nucleosome-protected DNA fragments using pooled MNase conditions show predominantly nucleosome-sized differences between the cell lines. While this was surprising, these differences occurred at developmentally important loci, suggesting we're seeing biologically significant changes in chromatin architecture. More specifically, we saw differences that corroborated earlier reports characterizing regulatory elements like enhancers as more open in active regions; we found enhancers in ESCs were more nucleosome-depleted than in differentiated cells. Using methodology that is standard in the field we saw no global changes, and this prompted us to consider optimizations to this methodology.

In Chapter two, we also addressed the question of what happens to the physical state of chromatin in the process of reprogramming to pluripotency by mapping the nucleosome

occupancy of iPSC. It is known that iPSC cells regain the global transcriptional signature, covalent histone tail modification patterns, and some macro characteristics of chromatin like protein turnover in the nucleus of ESCs, but whether the physical organization of local regions of chromatin were reprogrammed was still an open question. We find that the nucleosome occupancy profile of iPS cells is extremely similar to that of ES cells. ESC and iPSC have common RoDs with differentiated cells well outside of what could be expected by chance. Thus we conclude that the changes we see in this study are a hallmark of the pluripotent state.

As mentioned above, we hypothesized that MNase could be used in a more powerful way than the common methodology in the literature. When the DNA purified from a chromatin MNase digest is run on a gel it produces a characteristic ladder pattern, with 'rungs' in multiples of the size of DNA wrapped around a nucleosome, 150 base pairs. The extent of digestion impacts the number of rungs present. Published studies typically use one or two MNase conditions, and these conditions vary between studies (and possibly within a study). Thus, these studies profile a range of genomic populations. Some studies use DNA that comes from a digestion that produces only a single mononucleosome-sized band, while others use DNA from the mononucleosome-sized band of a ladder with 5 or more multi-nucleosome bands. These populations of DNA fragments represent different regions in the genome, however each study purports to be studying the location of nucleosomes in the genome.

We considered this problem in Chapter 2, and tried to address it by pooling of a range of MNase conditions. During this work it became obvious that a careful examination of separate MNase conditions would be useful to the field. In Chapter 3 we harness the activity of MNase

to produce a dataset more representative of the biological features it describes than previous datasets, and a new metric for clear interpretation and comparison of this data.

Using MNase titrations we find that open loci tend to have mononucleosome fragments liberated with low MNase conditions, and as digestion increases these fragments are reduced. Conversely, closed regions tend to have few mononucleosome fragments liberated under low digest conditions, but with increasing digestion more fragments are produced. Presumably fragments liberated under low MNase are more accessible by the enzyme, and the fragments produced with high MNase are less accessible, relatively speaking. Thus, we are able to assess whether a site is open or closed by our method. MNase titration data adds a new dimension to what is known about regulatory marks and proteins known to occur in open and closed regions. To produce a quantitative metric for looking at accessibility across the genome we developed a score for local regions based on the slope of averaged nucleosome occupancy, termed MACC. MACC scores also correlate with known features of chromatin, and can be derived for regions with little known about the local regulatory state. Further, MNase titration data and MACC scores can be derived for chromatin from any cell type making it a broadly applicable tool, and it is straightforward to perform. This new method that addresses the question of physical state of chromatin locally and globally is an important addition to chromatin biologists toolkit.

That being said, much more could be done to improve our understanding of how best to perform MNase titration digests. The high cost of sequencing prohibits a broad characterization of mammalian genome-wide data sets. If this were not the case, a thorough characterization of all variables potentially affecting MNase-seq that can be modulated in these experiments would be a worthy undertaking. The first characteristic of MNase-seq datasets I would assess is the

effect of differing levels of sequence coverage. The human genome has 3 billion base pairs; *D. melanogaster* has a genome 20 times smaller, and in our study the four titration points were sequenced at equal levels. The results we saw seemed 'cleaner' in *D. melanogaster*, allowing investigation in some of our analysis of smaller bin sizes. Short of sequencing 20 additional lanes of human sample we can only guess that the higher sequencing coverage in the *D. melanogaster* samples contributed at least in part to this observation.

Another notable difference between our work and the majority of MNase-seq experiments carried out in mammalian cell types is that we cross-linked our samples prior to digestion to stabilize the dynamic interplay of chromatin proteins. Small perturbations to chromatin can cause immediate alterations to regulatory chromatin landscape, so to perform meaningful comparisons we chose to remove the effect of the handling of cells prior to digestion. It would be interesting to see what effect this has on the downstream analysis and findings in comparison to native digestions.

It is both a plus and a minus that we investigate nucleosome occupancy in populations of cells. Averaging a population gives us an idea of the global features that are common to a population, but this averaging masks any dynamic features. It is widely known that even a cultured cell line is heterogeneous in its transcription, and cell cycle stage will also exert and effect on chromatin structure. Thus, it would be interesting to preform a cell number titration with set MNase conditions. Additionally investigation of individual cell cycle stages would be highly interesting. Library preparation may play some role in the results obtained from MNase-seq, in particular the range of fragments sizes can be affected by library preparation protocols. A comparison of SPRI clean up versus no clean up or even agarose gel clean up would tell us if

any bias was present in this procedure. Using adapter-free methods to prepare a library could allow a more sophisticated assessment of sub-nucleosomal sized protected fragments by allowing more small DNA fragments into the library. Of particular interest is the work by the Henikoff lab using salt fractionation. There is a long history of using different salt conditions to liberate different populations of chromatin, and joining these methods with the MNase titration method could refine our ability to characterize chromatin across the genome.

Our work and others' (whether in cross-linked or native MNase digestions) suggests that bound transcription factors may influence our findings. One additional line of investigation could include depletion of particular transcription factors using an antibody from MNase digested samples followed by sequencing the non-depleted fraction and the depleted fraction. This could shed light on the effect of that bound factor to a nucleosome occupancy map. A final experimental alteration that could be useful would be the addition of "spike-in" oligos of varying GC content for use as a measure of the effect of sequencing on the GC content of the sequence data produced.

In addition to examining and improving the MNase titration methodology it would be of interest to carry out MNase titration experiments in additional biological systems to see the extent of the ability of this method to pick up differences between conditions. Several examples that may be interesting include: Cells that exert a rapid response to stimulus, for example immune cells or heat shock, pioneer factor effect in a relevant cell model system, the effect of remodeler knockouts would be of particular interest given the effect of remodelers on nucleosome stability and accessibility, and of course further work on the system of development used in Chapter 2.

Taken together our work both advances our understanding of the physical state of chromatin in multiple mammalian cell lines and explores and improves upon methodology that has existed for decades. The clear diversity of nucleosome occupancy maps occurring after MNase titration informs the interpretation of future MNase mapping in the field, and it will be interesting to see this methodology applied to additional systems in the future.

Appendix 1 Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming

This paper is accepted and in preparation for publication. The uncorrected proofs are included here.

Jason A. West[*], **April Cook**[*], Burak H. Alver, Matthias Stadtfeld, Aimee Deaton, Konrad Hochedlinger, Peter J. Park, Michael Y. Tolstorukov, Robert E. Kingston (2014) Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming. Nature Communications, *in press*

nature COMMUNICATIONS

ARTICLE

Received 19 Jan 2014 | Accepted 16 Jul 2014 | Published xx xxx 2014

DOI: 10.1038/ncomms5719 OPEN

# Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming

Jason A. West[1,2,*,†], April Cook[1,2,*], Burak H. Alver[3], Matthias Stadtfeld[4], Aimee Deaton[1,2], Konrad Hochedlinger[5,6], Peter J. Park[3], Michael Y. Tolstorukov[1,*] & Robert E. Kingston[1,2]

Chromatin structure determines DNA accessibility. We compare nucleosome occupancy in mouse and human embryonic stem cells (ESCs), induced-pluripotent stem cells (iPSCs) and differentiated cell types using MNase-seq. To address variability inherent in this technique, we developed a bioinformatic approach to identify regions of difference (RoD) in nucleosome occupancy between pluripotent and somatic cells. Surprisingly, most chromatin remains unchanged; a majority of rearrangements appear to affect a single nucleosome. RoDs are enriched at genes and regulatory elements, including enhancers associated with pluripotency and differentiation. RoDs co-localize with binding sites of key developmental regulators, including the reprogramming factors Klf4, Oct4/Sox2 and c-Myc. Nucleosomal landscapes in ESC enhancers are extensively altered, exhibiting lower nucleosome occupancy in pluripotent cells than in somatic cells. Most changes are reset during reprogramming. We conclude that changes in nucleosome occupancy are a hallmark of cell differentiation and reprogramming and likely identify regulatory regions essential for these processes.



Q1 [1] Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [2] Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. [3] Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, USA. [4] The Helen L. and Martin S. Kimmel Center for Biology and Medicine, The Skirball Institute of Biomolecular Medicine, Department of Cell Biology, New York University School of Q2 Medicine, New York, New York 10016, USA. [5] Howard Hughes Medical Instituteand the Center for Regenerative Medicine, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [6] The Department of Stem Cell and Regenerative Biology, Harvard University, Cambridge, Massachusetts 02138, USA. * These authors contributed equally to this work. † Present address: Therapeutic Innovation Unit, Amgen, Inc., Cambridge, Massachusetts 02139, USA. Correspondence and requests for materials should be addressed to P.J.P. (email: peter_park@hms.harvard.edu) or to M.Y.T. (email: tolstorukov@molbio.mgh.harvard.edu) or to R.E.K. (email: kingston@molbio.mgh.harvard.edu).

Embryonic stem cells (ESCs) and induced-pluripotent stem cells (iPSCs) self-renew and differentiate into various cell types *in vitro* and *in vivo*. A complex network of genetic and epigenetic pathways regulates their self-renewal and differentiation, and the structural organization of chromatin play a prominent role in these processes. Prior studies have established multiple unique properties of pluripotent chromatin and its regulation, including macrostructural descriptions of ESC chromatin as relatively 'open' compared with lineage-committed cells[1–6]. The pluripotency factors Oct4, Sox2 and Nanog transcriptionally regulate and interact with specific chromatin-remodeling and histone-modifying complexes[7]. Reciprocally, multiple chromatin regulators, including complexes unique to ESCs, have been implicated in the maintenance of pluripotency, cellular differentiation and development[1–3,8–10].

The physical packaging of DNA into nucleosomes is a central determinant of DNA accessibility in both *cis* and *trans*. Nucleosomes consist of ~150 bp of DNA wrapped around a core histone octamer[11,12]. Nucleosome positioning on genomic DNA is dynamic and influences regulatory factor binding, which impacts processes ranging from gene regulation to DNA replication, recombination and repair[13,14]. Thus, characterizing changes in nucleosome occupancy should reveal important regulatory features in pluripotent cell biology, differentiation and reprogramming. Information on nucleosome location can be integrated with previous studies on covalent changes to chromatin (for example, DNA and histone methylation, histone acetylation) to improve our understanding of how chromatin dynamics contribute to pluripotency.

Techniques that map nucleosome positioning on the genome scale have illuminated the role of primary chromatin structure in the mammalian cell[15–22]. However, comparing the nucleosome profiles between different cell types still presents profound challenges. Observed nucleosome occupancy is sensitive to even slight variations in experimental conditions, such as the degree of chromatin fragmentation or chromatin isolation conditions[23,24]. This variability is hard to control and, as a result, changes in nucleosome occupancy and positioning associated with biological processes in mammals have been difficult to quantify. In particular, it is not clear if large-scale or local nucleosome profile rearrangements are prevalent upon cell fate change—how these rearrangements contribute to alterations in gene expression, and if nucleosome profiles are reset completely upon cell reprogramming.

Here we investigate nucleosome occupancy within pluripotent and somatic cells and identify regions of differences (RoD) between ESCs, iPSCs and somatic cells in both mouse and human. This analysis is facilitated by a novel data processing method developed for pairwise comparisons of nucleosome occupancy measured in different conditions and cell types. We report that the observed differences mostly do not appear to exceed the size of single nucleosomes, are enriched for motifs of transcription factors (TFs) that drive pluripotency and somatic cell reprogramming, and reside within gene regulatory regions, specifically at transcriptional start sites (TSSs) and enhancers of genes linked to pluripotency and differentiation. These findings reveal that localized changes in nucleosome occupancy at key regulatory regions, rather than large-scale rearrangements, may be sufficient to impact cell identity.

## Results

**Nucleosome mapping in pluripotent and somatic cells**. The results of this study are primarily based on the analysis of primary chromatin structure in three murine cell types: mouse ESCs, iPSCs derived from tail-tip-fibroblasts (iPSC-TTFs) and somatic TTFs. We also used somatic liver cells for validation purposes. All cells originated from the same isogenic mouse line and were previously characterized[25]. For each cell type, we created a genome-wide profile of nucleosome occupancy. To this end, we measured DNA protection patterns after chromatin digestion by micrococcal nuclease (MNase), building upon strategies previously developed by our group and others[15,17,20,26–29]. MNase selectively cleaves chromatin in linker DNA between nucleosomes, allowing a detailed description of nucleosome occupancy in a given cell population. The digestion fragments were size-selected and subjected to high-throughput sequencing, generating over 100 million mapped paired-end reads for each cell type. The average fragment size for each library was near the predicted mononucleosomal DNA length (~150 bp), and libraries showed high complexity with low percentages of repeats. We note that while the majority of the sequenced fragments likely represent nucleosome-associated DNA, some fragments may originate from loci protected by non-histone proteins, such as TFs[30]. Conversely, due to the preferential elimination of longer fragments during library preparation and sequencing, our data set may be depleted of the nucleosomes bound by larger complexes such as Pol II[31]. With these limitations in mind, we use the term nucleosome occupancy to characterize the number of digestion fragments at a given genomic position.

For comparison and validation of our results, we also used human ESCs, fibroblast-derived human iPSCs and differentiated fibroblasts (referred here as hESCs, hiPSCs and human fibroblasts, respectively). Here we emphasize the data from mouse cells, as we have a greater number of isogenic cell types for comparison and these data displayed higher reproducibility in our analyses. Importantly, the same trends were observed in the data derived from human samples (for more details, see Supplementary Material).

We first assessed the average nucleosome occupancy patterns at the TSSs for each cell type. As demonstrated previously[16,17,19,26], a nucleosome depleted region (NDR) flanked by well-positioned +1 and −1 nucleosomes (relative to the TSS) is a characteristic feature of the occupancy profiles averaged across all genes (Fig. 1). Indeed, we detected such a pattern across all samples (Fig. 1a). However, we also observed high variability in the magnitude of the nucleosome occupancy for ESCs and iPSCs, which show nearly identical gene expression patterns in both the mouse and human data (Fig. 1a and Supplementary Figs 1 and 4). Furthermore, such variation was observed even for biological replicates of the same cell type. This variability is not specific to our experimental protocol, as previous studies in mammalian genomes reported substantially different nucleosomal patterns at TSSs, ranging from an accumulation in tag counts greater than the surrounding regions to an apparent depletion in occupancy[16–19,22,32]. Thus, it likely originates from technical differences in experimental procedures, such as the extent of MNase digestion or chromatin isolation. This variability hinders direct comparisons of the nucleosome occupancy between cell types.

Among the characteristics of MNase-seq data that correlate with the degree of MNase digestion is the GC-content distribution of the sequenced fragments, which noticeably varied across all samples including biological replicates (Supplementary Fig. 2). The GC content of a population of MNase-digested DNA fragments can change with increasing or decreasing digestion levels[33]. This is in part due to MNase bias towards cutting AT-rich sequences, and in part due to preferential digestion of genomic regions with different accessibility and base composition. We expected GC-content distribution to be similar between replicates given our careful control of digestion

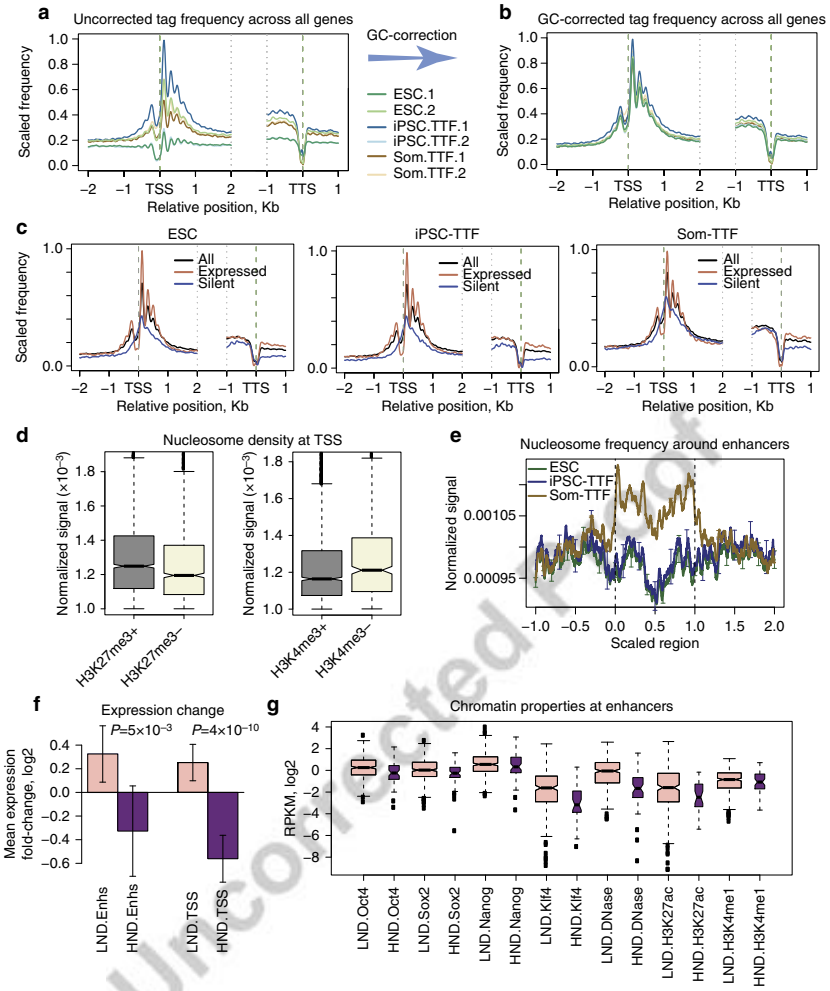





**Figure 1 | Comparison of nucleosome occupancy in mouse pluripotent and somatic cells.** (**a**) Nucleosome occupancy around transcription start and end sites computed for mouse ESCs, iPSCs and somatic tail-tip-fibroblasts (TTFs). We note that after normalizing the occupancy for the total number of tags in each library the profiles remain different, even between replicates of the same cell type. (**b**) The same profiles after normalization of the GC-content distribution in each sample with the target mean GC content of 50% (see Methods for more detail). (**c**) Comparison of the GC-normalized profiles for all genes and genes stratified by their expression status. (**d**) Boxplot showing nucleosome density distributions in TSS-proximal regions (±2 kb) stratified by the enrichment in H3K4me3 and K3K27me3 marks in ESCs. Notches at boxes provide reference to 95% confidence intervals. (**e**) Normalized nucleosome occupancy signal around scaled ESC enhancer regions computed for replicate sets in three cell types. (**f**) Comparison of gene expression and nucleosome occupancy changes. The two left bars show the expression changes computed for genes assigned to enhancers that have either lower (LND, pink) or higher (HND, purple) nucleosome occupancy in ESCs as compared with somatic TTFs; the two right bars depict the same for genes where nucleosome occupancy loss or gain occurs in the TSS-proximal regions. The 95% confidence intervals are shown with vertical arrows. (**g**) Comparison of the different chromatin properties (measured in ESCs[40]) for the LND and HND enhancers. As in **d** notches provide 95% confidence intervals.

conditions, DNA fragment selection and library preparation; however, we still observed variability. To address this issue, we implemented a step in our methodology that used the GC content of DNA sequence as a metric for normalization. Previously, nucleotide composition, including GC-content normalization, has been applied to the analysis of microarray and high-throughput sequencing data (ChIP-chip, chromatin immunoprecipitation sequencing (ChIP-seq) and DNA-seq and so on)[34–36]. Here we applied a concept used for ChIP-seq data[35] to the data produced by MNase digestion assays (Supplementary Fig. 3). We normalized GC content in each sample to 50%, which roughly corresponds to the average GC content in the TSS-proximal regions in the genome. The GC-content normalization markedly reduced variability across all TSS-proximal profiles in both

murine and human data (Fig. 1a,b and Supplementary Fig. 4A,B). Since TSS-proximal profiles are produced by averaging across large sets of genomic loci, they should be similar for samples demonstrating similar gene expression patterns, especially for replicates of the same cell type. To evaluate the extent of similarity, we computed correlation between nucleosome densities at TSSs in different samples (measured as average normalized frequency of fragments per kilobase of DNA) and observed increased correspondence between replicates of the same cell type upon GC normalization (Supplementary Fig. 5).

Nucleosomes differ in their properties including stability, accessibility and turnover rate, and the magnitude of the nucleosomal signal detected at TSSs in a particular study reflects how well nucleosomes of each type are profiled in a specific experimental setting. For example, using different salt concentrations during chromatin isolation results in different TSS-proximal profiles[24]. Similarly, different MNase digestion levels can produce different TSS-proximal profiles, each reflecting nucleosomal signal characteristic for given experimental conditions. Therefore, to further validate our results, we assessed another target GC content (48%, which represents the average GC content of our samples), confirming that our conclusions are not limited to a specific target GC content used for normalization (see Methods). Thus, we conclude that the GC normalization effectively reduces variability present in MNase-seq data and enables comparisons of nucleosome occupancy across different cell types. Equipped with this methodology, we proceeded to identifying changes in nucleosome occupancy in pluripotent and somatic cells.

**Nucleosome occupancy at regulatory loci varies in cell types.** We began by investigating differences in nucleosome organization at gene promoters and enhancers where we hypothesized it to play a role, and then extended the analysis to the whole genome. Using normalized MNase-seq data, we initially examined nucleosome occupancy of promoters in relation to both the transcriptional status of the associated gene and the covalent-histone modifications present. Consistent with previous reports, promoters of transcriptionally active genes showed an enhanced NDR as well as pronounced phasing of nucleosomes distal to the $+1$ and $-1$ nucleosomes, while promoters of transcriptionally silent genes lacked an NDR, demonstrating instead an occupancy signal indicative of a single nucleosome located approximately at the $+1$ nucleosome site (Fig. 1c)[16–19,22,32]. Furthermore, an increased NDR was observed in a cell-type-specific manner for genes that were upregulated in pluripotent cells (Supplementary Fig. 6). This effect was not pronounced for genes upregulated in somatic cells, suggesting that factors other than nucleosome rearrangement are responsible for silencing these genes in the pluripotent state.

Pluripotent cell promoters have been extensively characterized with regard to covalent-histone marks, including histone H3 lysine 4 trimethylation (H3K4me3) and histone H3 lysine 27 trimethylation (H3K27me3), which are associated with active and silent genes, respectively. Indeed, promoters classified by H3K4me3 and H3K27me3 enrichment exhibited nucleosome occupancy profiles characteristic for corresponding transcription status (Supplementary Fig. 7A–D). Comparing the average nucleosome occupancy at these promoters revealed decreased and increased occupancy levels for the promoters associated with H3K4me3 and H3K27me3 enrichment, respectively (Fig. 1d). This observation is consistent with increased nucleosome occupancy hindering transcription on average[37]. Interestingly, despite a lack of transcriptional activity at bivalent promoters (TSSs possessing both H3K4me3 and H3K27me3)[38], their nucleosomal profiles closely resembled those of transcriptionally active genes (Supplementary Fig. 7E). We note that most bivalent genes are associated with CpG islands, which may contribute to a chromatin structure that is poised for transcription activation during development[39].

Enhancers comprise another class of regulatory regions key for the pluripotent state. Here we used a recently-published set of enhancers associated with the pluripotency and reprogramming factors Oct4, Nanog and Sox2, including a subset of 'super-enhancers' that are unusually large and impart hyper-regulatory functions in ESCs[40,41]. The set comprises 8,794 enhancers, 231 of which are super-enhancers. Comparison of the nucleosome occupancy profiles around scaled ESC enhancers in somatic and pluripotent cells revealed that on an average the occupancy was lower in pluripotent cells (Fig. 1e), which is consistent with these regions being more accessible to regulatory proteins in pluripotent cells. The same trend was observed in human MNase-seq data for hESCs, hiPSCs and differentiated human fibroblasts (Supplementary Fig. 8A).

For a more detailed analysis, we divided all enhancers into two groups, those having significantly lower nucleosome density (LND) or higher nucleosome density (HND) in ESCs when compared with somatic TTFs (significance based on the variability of the nucleosome density in the replicates; *t*-test, *P*-value threshold 0.05). Consistent with the results described above, the LND group comprised 353 enhancers (23 of which were super-enhancers), while the HLD group comprised only 60 enhancers (one of which was a super-enhancer). When all the TSS-proximal regions were similarly divided into LND and HND groups for comparison, the corresponding counts were 558 and 341, thus resulting in considerably less skewed group counts than those detected for enhancers. We note that more than a twofold skew in the numbers of LND and HND enhancers was also present when the comparison included all enhancers rather than being limited only to those showing statistically significant differences (Supplementary Fig. 9A).

The expression of genes associated with enhancers from the LND and HND groups significantly differed in ESCs and somatic cells (Fig. 1f, $P = 5 \times 10^{-3}$, *t*-test; Supplementary Fig. 9B), with the genes associated with LND enhancers showing higher expression than the genes associated with HND enhancers. This difference was approximately the same in magnitude as that observed for the LND and HND promoters.

To further investigate how nucleosome occupancy at enhancers correlates with other features of chromatin organization, we used published data on chromatin structure and TF binding in ESCs[40]. Enhancers with LND were more likely to be bound by pluripotent TFs, exhibited active chromatin marks and were associated with stronger DNase I signal when compared with enhancers from the HND group (Fig. 1g). This rearrangement of the nucleosome landscape at enhancers might be a key determinant in pluripotency and differentiation, with lower nucleosome occupancy correlating with stronger enhancer activity in pluripotent cells. We conclude that the rearrangement of the nucleosome landscape at regulatory regions correlates with changes in other chromatin signatures and gene expression in a cell-type-specific manner, and that active enhancers show lower levels of nucleosome occupancy in pluripotent cells.

**Punctate changes at regulatory regions discern cell types.** We next sought to identify all RoD in the nucleosome occupancy profiles of ESCs, iPSCs and somatic cells on a genome scale, regardless of their location relative to the annotated DNA elements. Nucleosome organization is likely to undergo rearrangement as cells change fate, and visual inspection of the

nucleosome occupancy profiles indeed revealed such changes (Supplementary Fig. 10). However, little is known about nucleosome occupancy changes on the genomic scale, including their significance, prevalence, size and distribution, in part due to the challenges inherent in mapping these differences in mammalian cells. We applied a novel approach comparing the frequency of digestion fragments in 150-bp bins to scan the genome and generate *P*-value profiles, describing the significance of nucleosome occupancy differences between cell types (Fig. 2a). We used a false discovery rate (FDR) threshold to identify sets of significant RoDs for each pairwise cell-type comparison; we note that since this algorithm is not focused on stable nucleosome positions, it is suitable for detection of RoDs of any size (see Methods for details). To rule out the possibility that RoD detection is driven by an outlier replicate, we confirmed that the direction of change in nucleosome occupancy at RoDs is the same in all pairwise comparisons of the replicates (Supplementary Fig. 11).

Our approach is further illustrated in Fig. 2b, showing the promoter of Oct4 gene. This gene has a nucleosome occupancy

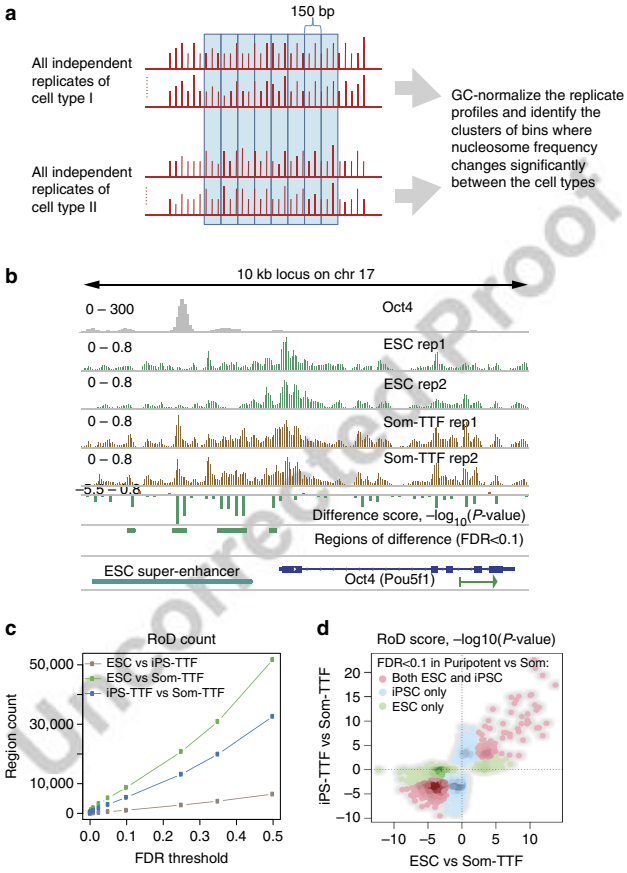



**Figure 2 | Identification and characterization of regions of difference (RoDs) in nucleosome profiles between murine pluripotent and somatic cell types.** (**a**) Schematic illustration of the method used for RoD identification. In short, sequenced tag frequencies in all replicates of the compared cell types (red) were binned along the genomic coordinate (blue) and the clusters of the bins where tag frequencies were significantly different were retained for further analysis (see Methods for detail). (**b**) Normalized nucleosome occupancy in the promoter of the Oct4 (Pou5f1) gene for two independent ESC lines and isolates of somatic TTFs. The computed difference score and identified RoDs are shown as green vertical and horizontal bars, respectively, below the occupancy tracks. The sign of the difference score indicates whether nucleosome occupancy was gained (positive score) or lost (negative score) in the 'ESC versus somatic TTF' comparison. The green arrow next to the gene name indicates direction of transcription. (**c**) Counts of the RoDs identified with different FDR thresholds (FDR = 0.1 was selected to compose the representative RoD sets for the downstream analyses). (**d**) Correlation between difference scores of the RoDs identified in comparisons of ESCs versus somatic TTFs and iPS-TTFs versus somatic TTFs. Only the bins that meet the FDR threshold of 0.1 at least in one comparison were taken for this analysis. Red dots represent bins that meet the selected FDR threshold in both comparisons; blue and green dots represent bins that meet the FDR threshold only in the 'iPSC versus somatic TTF' set or only 'ESC versus somatic TTF' set, respectively. We note that the sign of the score is maintained across the sets (that is, the bins that have positive (negative) scores in one pairwise cell-type comparison have the same score signs in another pairwise cell-type comparison), which is indicative of good correspondence between the loci of nucleosome occupancy variation in ESCs and iPS-TTFs.

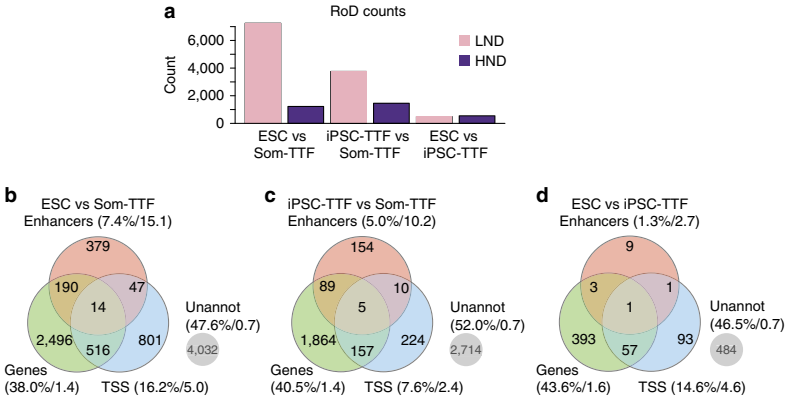



**Figure 3 | Occurrences of the regions of difference (RoDs) identified in pairwise comparisons of mouse cell types.** (**a**) Comparison of the counts of RoDs with lower (pink) and higher (purple) levels of nucleosome occupancy in the pluripotent cell types relative to somatic TTFs (first two bar groups) and in ESCs relative to iPSCs (last bar group). (**b**–**d**) Occurrences of the identified RoDs in the different regions of the genome for pairwise comparisons of ESCs versus somatic TTFs (**b**) iPS-TTFs versus somatic TTFs (**c**), and ESCs versus iPS-TTFs (**d**). Genes are defined according to USCS annotation for mm9 genome, TSS-proximal regions comprise ±2 kb around gene starts, and ESC enhancer coordinates were taken from a recent publication[40]. The numbers inside the circles represent counts of RoDs in corresponding regions. The numbers next to the region name represent the percentage of the RoD occurrences in this region to the total RoD count and the enrichment of this percentage over the expected value based on the region size in the genome. We note that the evaluated regions can overlap and therefore the sum of the percentages is not equal to 100%. This figure only includes RoDs meeting a FDR = 0.1.

pattern characteristic of an expressed gene in pluripotent cells, with an NDR at the TSS flanked by regions of high nucleosome occupancy. Somatic cells, which do not express Oct4, lack the NDR at the Oct4 TSS and show overall higher levels of nucleosome occupancy in the promoter region. Our approach was able to detect these changes and identify the RoD overlapping an Oct4 binding site important for gene upregulation in ESCs[40]. GC normalization, one of the features that distinguishes our approach from earlier algorithms[42], facilitated the identification of RoDs by reducing variability between replicates and allowed identification of more RoDs by ~45% in the comparison of ESCs and somatic TTFs, including those at the Oct4 locus (Supplementary Fig. 12).

To evaluate the extent to which somatic cell reprogramming resets the chromatin structure in iPSCs, we compared the numbers of RoDs identified between pluripotent and somatic cell types with those detected between ESCs and iPSCs. As the number of detected RoDs is a function of the selected significance threshold, we analyzed RoD counts for a series of FDR thresholds. We consistently identified more RoDs in pluripotent versus somatic cell comparisons than comparisons of two independent pluripotent cell lines (Fig. 2c). For instance, at FDR = 0.1, we identified over 8,000 RoDs when ESCs were compared with somatic TTFs, and over 5,000 RoDs when iPSCs were compared with somatic TTFs. For the ESCs and iPSC comparison, 1,041 RoDs were identified, which is five to eightfold lower than the number of RoDs identified in any pluripotent versus somatic cell comparison. We note that the transcriptional profiles of ESCs and iPSCs were very similar (Supplementary Fig. 1F), with <50 genes demonstrating significant changes in expression (see Methods for details on calling differentially expressed genes), which is consistent with the low number of RoDs detected when comparing these cell types.

iPSCs could more closely resemble their cell of origin rather than ESCs with regard to nucleosome placing. However, based on previous work, ESCs and iPSCs are functionally equivalent and very similar at the molecular level (reviewed in ref. 43), and thus one would anticipate a high degree of similarity between iPSCs and ESCs in nucleosomal occupancy profiles. Indeed, the differences in nucleosome organization observed in the comparisons of somatic cells to ESCs correlate with the differences detected in comparisons with iPSCs (Fig. 2d). For instance, all the regions determined for a selected FDR threshold in ESCs exhibit the same directional change in the iPSC comparison, and vice versa (green and blue dots in Fig. 2d). These observations were further confirmed in hESC, hiPSC and human fibroblast comparisons (Supplementary Fig. 4C,D).

We also examined two basic characteristics of RoDs: their size distributions and the direction of nucleosome occupancy change. Surprisingly, while nucleosomes with altered occupancy might cluster, a vast majority of RoDs appeared to be 150 bp in size (>95% in both the mouse and human data). A small percentage (<1%) of RoDs were several kilobases in length, but no regions >0 kb were observed (Supplementary Fig. 13). We note that the resolution of our approach as well as the smallest RoD size that can be reported is 150 bp, which is the size of the bins used for this analysis. Therefore, we cannot distinguish between changes occurring on mononucleosomal versus subnucleosomal scales. Our technique, however, would detect changes occurring on larger scales as those spanning multiple adjacent bins. Low count of RoDs exceeding 150 bp allows us to conclude that such large-scale changes in nucleosome occupancy are infrequent, suggesting tight control of chromatin structure at the level of single nucleosomes. When directionality of the occupancy change was considered, the majority of the RoDs identified between pluripotent and somatic cells showed an increase in nucleosome signal in the differentiated cells (Fig. 3a, see Supplementary Fig. 4E for human data). This supports the hypothesis that pluripotent cells have relatively open chromatin, as one criterion for open chromatin would be lower nucleosome occupancy. The RoDs identified between ESCs and iPSCs showed little bias in nucleosome occupancy change direction, suggesting the absence of a dominant trend distinguishing the chromatin structure in these cells.

Thus our analysis revealed mostly punctate differences in nucleosome occupancy between pluripotent and somatic cells. These loci are predominantly associated with lower nucleosome occupancy in pluripotent cells. Overall, ESCs and iPSCs display a high degree of similarity in nucleosomal signal, providing evidence that somatic cell reprogramming resets nucleosome positioning to a pluripotent state[44]. We next sought to more fully characterized RoD locations, as these regions are likely regulatory sites involved in pluripotency and reprogramming.

**RoDs are enriched at regulatory regions active in ESCs.** Our analysis showed that ~40% of RoDs are at gene regions annotated in the mouse genome (Fig. 3b–d, see also Fig. 4a, Supplementary Fig. 8B), which is significantly more than expected for a randomized distribution ($P = 10^{-12}$, see Methods for details on significance estimation). Around genes, TSS-proximal regions are specifically enriched in RoDs (Fig. 4c, blue line), including the promoters of genes associated with pluripotency and transcription activation (as exemplified by Oct4 in Fig. 2b). Indeed, in pluripotent versus somatic cell comparisons, between 7 and 16% of RoDs occur at TSSs, and these are enriched 2.4 to 5 fold over the genome average (Fig. 3b,c). In addition to genes and their promoters, pluripotency-associated enhancers exhibited significant enrichment in RoDs (Fig. 4c, orange line, and Supplementary Fig. 8C). To our surprise, enhancers demonstrated differences with the same or greater magnitude as TSSs. In pluripotent versus somatic cell comparisons, between 5 and 7.4% of RoDs occurred at ESC-defined enhancers, which corresponds to a 10- to 15-fold enrichment over the genome-wide occurrence of these enhancers (Fig. 3b,c). 'Super-enhancers'—large enhancer regions associated with a high density of regulatory protein binding[40]—showed even stronger enrichment in RoDs (Fig. 4c, red line). As an additional validation of this result, we identified RoDs between ESCs and another somatic cell type, mouse liver. This set of RoDs was also skewed towards LND enhancers in ESCs and showed enrichment at TSSs and ESC enhancers (Supplementary Fig. 14), confirming that these effects are not specific to the somatic cell type to which ESCs are compared.

To further quantify the overlap between RoDs and these regulatory regions, we computed the percentage of enhancers and TSSs harbouring RoDs. We note that actual values of such an overlap would depend on the sequencing depth achieved in a particular study (that is, statistical power to identify RoDs and enhancers) and the significance threshold used to call RoDs. Under the threshold used in this study, we found that 7% of 'regular' enhancers and 39% of super-enhancers bear at least one RoD, which represents a significant overlap as compared with the expected value for randomized RoD distributions ($P = 10^{-11}$, see Methods). A similar fraction of TSS-proximal regions (6%) harbour RoDs, which reinforces the importance of chromatin structure and its regulation at enhancers in pluripotent and somatic cells. While most enhancers harbour only one or no RoDs, super-enhancers are often associated with multiple RoDs. An example of such a super-enhancer is given in Fig. 4b, where up to nine RoDs, all from the LND group, can be detected.

**RoDs are enriched in binding motifs of reprogramming TFs.** Given that the detected RoDs are small in size (~150 bp) and enriched at regulatory sites, one could hypothesize that they are associated with regulatory protein binding that displaces a single nucleosome. For instance, regions associated with binding of TF involved in cell differentiation were reported to have lower nucleosome occupancy in the corresponding somatic cell type[21]. Here we focused on the regions with lower nucleosome occupancy in pluripotent cells (LND RoDs) and analyzed them for the presence of sequence motifs to identify potential regulatory factors. We found that mouse LND RoDs identified in comparison of ESCs and somatic cells were enriched in motifs of TFs associated with reprogramming and pluripotency, including Klf4, c-Myc, Oct4 and Stat3 (Fig. 5a, Supplementary Fig. 15). As Oct4 and Sox2 act as heterodimers in pluripotent cells[45–47], we conclude that our analysis identifies potential sites of functional binding for all four Yamanaka reprogramming factors. The Stat3 motif is also highly enriched in these RoDs, and Stat3 is required and sufficient for the self-renewal of mouse ESCs[48]. Performing a *de novo* motif search with a random set of genomic sequences mimicking the RoD set did not reveal motifs for the Yamanaka factors (with the selected significance threshold of $E$-value $= 10^{-5}$). We note that many of the factors associated with the motifs identified within the RoDs also bind enhancers in pluripotent cells and, furthermore, their binding is often used to define enhancers in pluripotent cells[40,47].

Protein binding was previously shown to order nucleosomes on a scale larger than the 150 bp observed for most of the RoDs in our analysis[18,49,50]. We therefore examined how TF binding may affect nucleosome occupancy beyond the RoD boundaries in different cell types. To this end, we compared the nucleosome profiles around TF binding motifs in each cell type. Our results show that such TF-proximal nucleosome profiles exhibit unique properties depending on the TF considered. For the Oct4 motif, we observed clear nucleosome phasing emanating away from the Oct4 binding site in pluripotent cells but not in somatic TTFs, which lack Oct4 expression (Fig. 5b). Conversely, for a TF specific for differentiated cells, Hnf4a, we observed phasing in somatic but not in pluripotent cells (Fig. 5c). For c-Myc/Max (a TF that is expressed in ESCs, iPSCs and somatic TTFs), we observed nucleosome phasing in all samples (Fig. 5d). Interestingly, there is a shift in phasing with c-Myc/Max in pluripotent and somatic cells, which may indicate preferential binding of this TF to different genomic regions in these cell types. Together, these data support that local changes in nucleosome occupancy are formed around TF binding sites and suggest that the cell-specific TF expression and binding helps to establish the unique chromatin context for a given cell type[26,51,52].

To further validate that RoDs reflect TF binding sites, we investigated the enrichment of ChIP-Seq signal at these loci, using data on pluripotency-associated TF binding from an independent study[40]. Our results revealed several-fold enrichment in Oct4, Sox2 and Nanog signal at LND RoDs, while no such enrichment was detected for HND RoDs (Fig. 6a–d). In addition, the profile of H3K4me3, also based on ESC data, showed a clear drop at the center of LND RoDs, which is consistent with nucleosome depletion. These findings highlight a possible role for TF binding in the rearrangement of nucleosomal landscape and suggest that different factors are responsible for the emergence of LND and HND RoDs.

Overall, our results revealed that the differences in nucleosome occupancy profiles in pluripotent and somatic cells mostly appear as punctate changes at individual loci. These differences tend to cluster at regulatory regions that control gene expression, including promoters and enhancers of developmentally regulated genes, indicating their functional importance for determining the regulation of cell status. We conclude that these are not wholesale changes in nucleosome positioning between pluripotent and somatic lineages, but rather specific changes whose location implies a key role in the transition between cell states.

## Discussion

The objective of this study was to determine the nature of changes that occur in nucleosome occupancy profiles upon transition

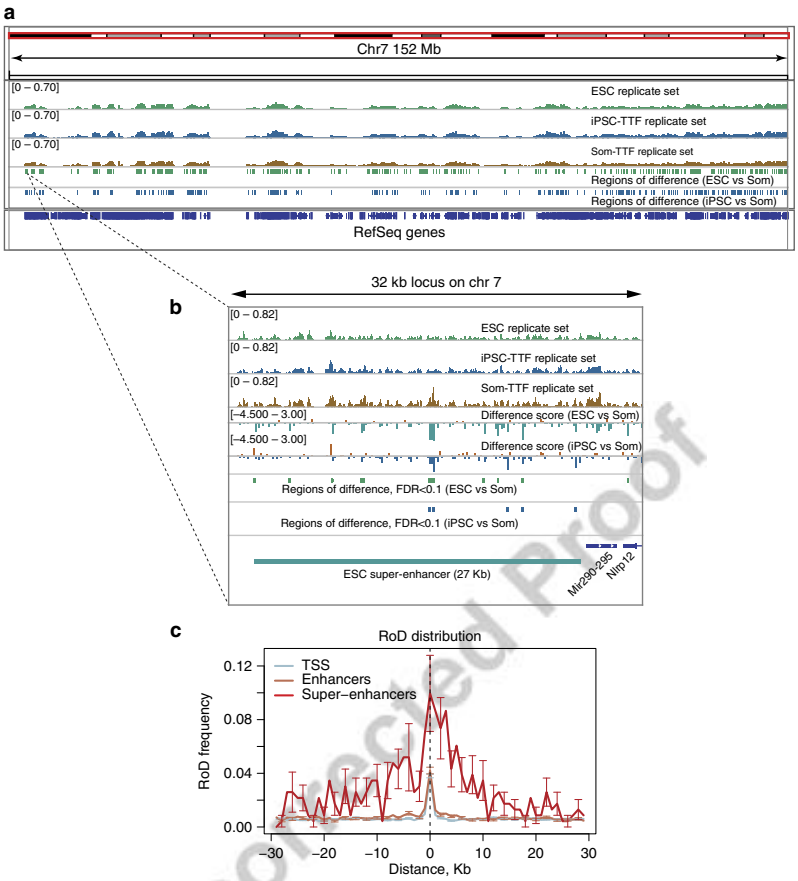



**Figure 4 | Distribution of the regions of difference (RoDs) detected in nucleosome occupancy profiles relative to annotated regions in the mouse genome.** (**a**) Chromosome wide snapshot of the normalized nucleosome occupancy and RoD occurrence. (**b**) Nucleosome occupancy at one of the super-enhancers identified in Whyte *et al.*[40] shown as an example of multiple RoDs present in this class of enhancers. (**c**) The RoD frequencies in the regions encompassing TSSs and enhancers identified in ESCs[40]. The 95% confidence intervals are shown with the vertical arrows. The confidence intervals were estimated based on the variability of the frequency values in individual profiles used for averaging.



between pluripotent and somatic cells. To address this question, we used MNase digestion assays as the primary tool. We note that while the extent to which MNase-associated bias affects the determination of nucleosome positioning is still debated[33,53], the design of our study, which involves an additional step for bias correction and focuses on pairwise comparison of the occupancy profiles, minimizes the possibility of artifacts.

One can expect that a dramatic change in cell identity, such as that occurring during somatic cell reprogramming or differentiation of pluripotent cells, would be accompanied by large-scale changes in primary chromatin structure. To our surprise, we detected only a handful of RoDs larger than 1 kb in size. At the same time, we observed a number of important features in the reorganization of nucleosomal landscapes associated with pluripotency. Our main conclusions are that changes in nucleosome occupancy largely do not exceed mononucleosomal size, co-localize with binding sites of pluripotency and reprogramming associated proteins, generally have reduced levels of nucleosome occupancy in pluripotent cells, and are enriched at enhancers, promoters and within genes (Figs 3b–d and 6e). Comparisons of different classes of regulatory regions revealed that RoDs at enhancers and especially at super-enhancers are at least as prevalent as those at TSSs, underscoring the importance of these regions in determining cell state[40,41,54].

Another central conclusion is that fully reprogrammed and characterized iPSCs[28,55] demonstrate nucleosome occupancy patterns similar to those in blastocyst-derived ESCs, with eightfold fewer RoDs detected between ESCs and iPSCs than between ESCs and somatic TTFs. Importantly, the nucleosome configuration at enhancers in iPSCs is similar to that in ESCs, while it is considerably different from that in fibroblasts. In addition, the RoDs identified between pluripotent and somatic cells contained binding motifs for the Yamanaka reprogramming factors as well as other key pluripotency factors, suggesting that the nucleosome occupancy changes overlap with the regulatory regions that are important for cell identity. Chromatin structure

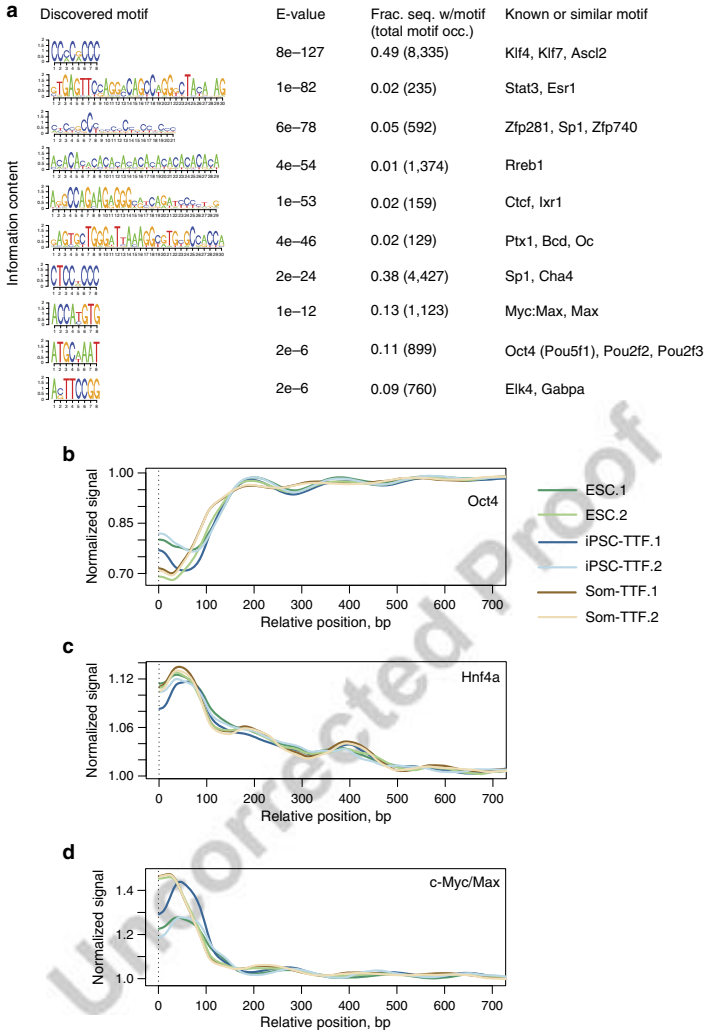





**Figure 5 | Relation between local nucleosome organization and the presence of TF binding motifs.** (**a**) Sequence motifs found in *de novo* enrichment analysis of the RoDs associated with lower nucleosome occupancy levels in ESCs as compared with somatic TTF cells. Corresponding *E*-values are indicated for each motif as well as the fractions of the test sequences with the motifs and total occurrences of the motifs in the sequence set, computed for 85% identity threshold. The last column lists the TFs associated with similar motifs are indicated. Motifs with no known associated protein and those <8 bp in length are not shown (see Supplementary Fig. 15 for a complete list of the identified motifs). (**b**–**d**) Distribution of nucleosome occupancy around the motifs of selected TFs, (**b**) Oct4, (**c**) Hnf4a and (**d**) c-Myc/Max. The occupancy was averaged over all motifs identified in the mouse genome with the selected FDR threshold and the plot was symmetrized relative to the motif center.

in general, and nucleosome occupancy in particular, could represent an additional and fundamental level of epigenetic memory that must be reset for proper somatic cell reprogramming[54,56].

Our analysis supports, from a distinct angle, the previously-reported observation that pluripotent cells have more 'open' chromatin compared with somatic cells. ChIP-seq on H3K9me3 and H3K27me3 suggested that these heterochromatic marks cover over three times more of the genome in differentiated cells when compared with ESCs[57]. In addition, the nuclei of pluripotent cells have macroscopic characteristics of less-condensed chromatin, histone turnover appears more dynamic in pluripotent cells, and regulatory regions show enrichment in histone variants and covalent modifications that are characteristic of open chromatin[4,58]. Here we observe that a majority of the detected RoDs are associated with lower nucleosome occupancy in pluripotent cells when compared with somatic cells. The lower nucleosome occupancy in pluripotent cells correlates with function, since it is predominantly observed at active chromatin

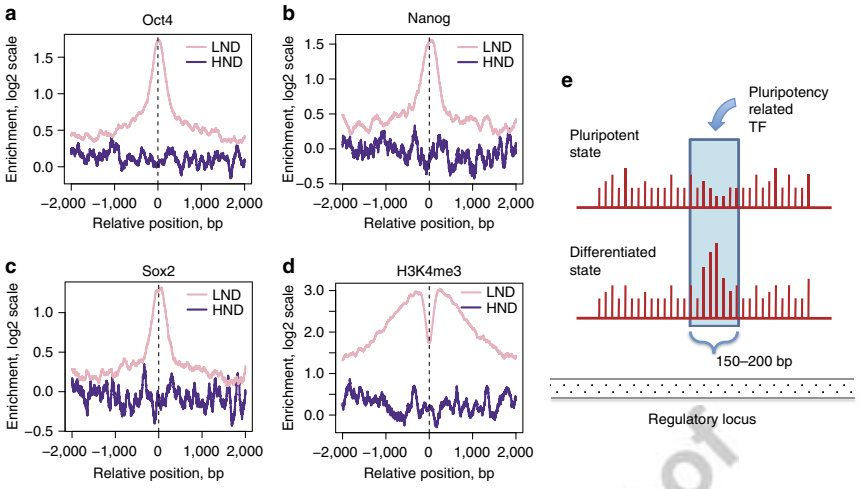



**Figure 6 | TF binding at the sites of nucleosome rearrangement.** Enrichment profiles (ChIP over WCE input) computed in the RoD proximal regions for (**a**) Oct4, (**b**) Nanog, (**c**) Sox2 and (**d**) H3K4me3 mark. Two classes of RoDs are considered separately: LND (light pink) and HND (purple). (**e**) Schematic summary of the observations reported in this paper. While nucleosome occupancy profiles (red vertical bars) remain similar between the pluripotent and differentiated states, there are punctate regions of difference (marked by the light blue rectangle) characterized by lower nucleosome occupancy in the pluripotent state. Majority of these regions do not exceed the size of a single nucleosome. They are enriched in binding motifs of pluripotency-related TFs and occur within regulatory regions, such as gene promoters and enhancers.

regions, including ESC-specific enhancers and promoters of genes upregulated in ESCs and iPSCs. We conclude that the more permissive chromatin configuration in pluripotent cells is enabled not only through reduction of the heterochromatic regions but also through local changes in the nucleosomal landscapes of euchromatic regions.

While most of RoDs do not appear to exceed the size of a single nucleosome, we note that protein binding may induce larger-scale rearrangement of chromatin, such as the increased nucleosome phasing observed in Fig. 5b–d. However, deeper sequencing and a larger number of replicates would be required to identify a 'complete' set of RoDs, which would include such changes at individual loci. In combination with protein-binding motif information, our current approach can be used for simultaneous identification of nucleosome rearrangement and differential binding for a range of TFs in one assay, when such data are available. This approach could be further enhanced by analyzing the digested fragments of subnucleosomal sizes and/or by using multiple levels of digestion for the same sample to preferentially profile genomic loci of different accessibility[30,59]. Such a comprehensive approach would help us better understand how changes in chromatin organization translate into changes in gene expression and cell identity.

## Methods

### Experimental procedures

*Cell culture.* Mouse ESCs and iPSCs were maintained on mouse embryonic fibroblast feeder layers (Specialty Media) in DMEM containing 15% heat-inactivated foetal bovine serum (Hyclone) supplemented with 1,000 U ml$^{-1}$ leukaemia Inhibitory Factor (Chemicon). The following mouse cell lines were used in these studies: A5 ESCs (ESC.1), A6 ESCs (ESC.2), A4 iPSCs (iPS.TTF.1) and A5 iPSCs (iPS.TTF.2). All isogenic lines were created from mice containing the stable integration of doxycycline-inducible reprogramming factors (Oct4, Sox2, Klf4 and c-Myc). All experiments were initiated with cell lines between passage 15 and 22. Primary TTFs and liver were obtained as secondary derivates from B6/129 neonatal mice aged between 7 and 14 days postpartum. These mice and cell lines have been functionally characterized and were previously reported[25].

Human ESCs and iPSCs were maintained on Geltrex (Life Technologies) in mTeSR1 (Stem Cell Technologies). H1-OGN ESCs[60] and iPSCs[28] were functionally characterized previously[28,60]. These cells exhibited the expected *in vitro* molecular and functional properties of human pluripotent cells in our hands, but showed low to no OCT4-GFP reporter expression. Experiments were carried out with H1-OGN ESCs between passage 76 and 77 and iPSCs between passage 14 and 17. Differentiated fibroblasts were made from H1-OGN ESCs and were used between passages 7 and 14.

*Chromatin digestion with MNase.* Each murine cell type was expanded to $\sim 3 \times 10^7$ cells and pretreated with mild detergents (0.2% Tween-20 and 0.2% Triton X-100) for 5 min followed by a 1.1% formaldehyde treatment for 10 min to preserve chromatin structure. Nuclei were then prepared from the cross-linked cells and the chromatin treated with three MNase concentrations for 15 min at room temperature (RT). A range of digestion conditions was employed to sample both hyper- and hypo-accessible chromatin regions to MNase digestion. Cross-links were then reversed for 16 h at 55 °C along with proteinase K digestion and DNA harvested via phenol–chloroform. Samples were then run on 1% agarose gels and the resulting mononucleosomal DNA fragments ($\sim$150 bp) were gel purified, pooled and prepared for sequencing on an Illumina HiSeq instrument.

Human cells were expanded to $\sim 1 \times 10^8$ cells and cross-linked with 1.1% formaldehyde for 10 min at RT. Nuclei were isolated and treated with a range of four MNase concentrations for 15 min at RT. Cross-link reversal was performed at 65 °C for at least 16 h followed by an RNase and subsequent proteinase K digestion. DNA was purified by phenol–chloroform extraction. Ampure SPRI beads (Beckman Coulter) were used in a double size selection with ratios of 0.7 × and 1.7 × to obtain a range of fragment sizes from $\sim$100 to 1,000 bp. The resulting sample contains a majority of mononucleosomal fragments with some smaller and di-nucleosome-sized fragments with high reproducibility. The resulting fragments from each MNase concentration in the range were prepared individually for barcoded sequencing on an Illumina HiSeq instrument. Mapped read from all concentration were subsequently pooled for analysis.

*Illumina HiSeq library preparation and sequencing.* Mononucleosome DNA (1 μg) was used for library preparation, with limited number of PCR amplification rounds[61], and genomic alignments of paired-end 50 bp reads were performed using Bowtie[62] followed by further tag processing and filtering with the SPP workflow[28]. All alignments and annotations used the mouse genome assembly mm9 and the human genome assembly hg19.

*Transcriptional profiling.* RNA samples from each cell line were purified using TRIZOL (Invitrogen), and double-stranded complementary DNA (cDNA) was generated using the SuperScript double-stranded cDNA kit (Invitrogen). Samples

were then submitted to Roche NimbleGen for subsequent hybridization and downstream processing using the NimbleGen 12 × 135 k mouse gene expression array platform, which assays 44,170 target genes with three separate 60mer probes per transcript. Biological replicates were performed for all cell lines.

### Bioinformatic and statistical data analysis

*Sequencing data preprocessing and initial analysis.* See Supplementary Table 1 for the number of tags and the insert size for each sample. Sequenced 50-bp paired-end tags were mapped to the mouse (mm9) or human genome (hg19) for the corresponding cell types using the Bowtie aligner v. 0.12.7 (ref. 62). Only uniquely mapped tags with no more than two mismatches in the first 28 bp of the tag were retained. Genomic positions with the number of mapped tags above the significance threshold of $z$-score = 7 were identified as anomalous, and the tags mapped to such positions were discarded. The coordinates of the genes were taken according to the annotations for mm9 and hg19 versions of the mouse and human genomes, respectively. The gene proximal profiles were calculated and plotted as described previously[29,63].

*GC-content normalization.* The GC-correction procedure applied in this study is illustrated in Supplementary Fig. 3. The correction coefficient for each read was computed in such a way that the resulting genome-wide distributions of GC content become similar to the target GC-content distribution (Gaussian distribution with mean GC = 50 and 48% and variance = 7.5%). Specifically, all reads were stratified according to the GC content of the regions ± 100 bp around mapping location of the pair-end read centres and the correction coefficients were computed as ratios of the histograms corresponding to experimental and theoretical GC-content distributions with 1% GC-content step. The coefficients were applied to the tag frequencies at each genomic position with non-zero tag counts. For the purpose of RoD identification, the corrected tag frequencies were rounded to the closest integer. The value of GC = 50% was used to obtain main results in the study, and GC = 48% was used for validation purposes to confirm that the same trends can be observed in downstream analyses with other target GC-content values (Supplementary Fig. 16).

*Detecting RoD in nucleosome occupancy.* $P$-values of the differences were estimated for the frequency of reads summarized within 150-bp non-overlapping bins. The $P$-value calculation was based on the negative binomial distribution, with variance and mean estimated based on the replicate profiles produced for each cell type, as implemented in the R package DESeq[64]. Default parameters of DESeq were used for computations. To account for local context of nucleosome occupancy, the estimation of significance of the nucleosome occupancy changes within bins was performed independently in 25 kb segments with a 12.5 kb step, hence generating two significance values for each bin. The more conservative estimate was retained for further analysis. The bins exhibiting significant changes in tag frequency between the samples separated by > 100 bp were merged to form RoD. Coordinates of the identified RoDs are provided as Supplementary Files.

*Estimation of statistical significance.* Significance estimations were performed using R (http://www.r-project.org). Abundances of RoDs in genomic regions were compared with the corresponding values obtained for the randomized RoD distributions in mappable regions of the genome using non-parametric Wilcoxon test (as implemented in the function 'wilcox.test' from the package 'stats'). Only the regions of the genome that had non-zero tag counts were used in randomization (at least 1,000 randomizations were performed in each case).

*Gene expression data processing.* Gene expression data for mouse cells were generated using the NimbleGen expression micorarrays (Roche NimbleGen, Inc., Madison, WI). Microarray data provided by NimbleGen were background-corrected and normalized between the arrays using the Robust Multichip Average package. Fold-change and statistical significance were estimated for the log2 expression values of each gene based on the data for individual replicates within each replicate set. The genes with at least twofold change in expression and meeting a $P$-value threshold of 0.05 were identified as differentially expressed.

*Motif analysis.* Motif analysis was performed using web-base service MEME-ChIP[65]. Motifs that are at least 6 bp in length identified with an $E$-value threshold of $10^{-5}$ were reported. Both palindromic and non-palindromic motifs were allowed. The motifs found in the test sequences were matched against JASPAR (CORE-2009) or UniPROBE (mouse) databases to identify similarity with known protein motifs using tools implemented in MEME-ChIP with default parameters. Calculation of motif occurrences in test sequences and sequence logo generation were performed using the Bioconductor packages Biostrings and seqLogo, respectively (http://www.bioconductor.org).

## References


1. Gao, X. *et al.* ES cell pluripotency and germ-layer formation require the SWI/SNF chromatin remodeling component BAF250a. *Proc. Natl Acad. Sci. USA* **105,** 6656–6661 (2008).
2. Ho, L. *et al.* An embryonic stem cell chromatin remodeling complex, esBAF, is essential for embryonic stem cell self-renewal and pluripotency. *Proc. Natl Acad. Sci. USA* **106,** 5181–5186 (2009).
3. Loh, Y. H., Zhang, W., Chen, X., George, J. & Ng, H. H. Jmjd1a and Jmjd2c histone H3 Lys 9 demethylases regulate self-renewal in embryonic stem cells. *Genes Dev.* **21,** 2545–2557 (2007).
4. Meshorer, E. *et al.* Hyperdynamic plasticity of chromatin proteins in pluripotent embryonic stem cells. *Dev. Cell* **10,** 105–116 (2006).
5. Fussner, E. *et al.* Constitutive heterochromatin reorganization during somatic cell reprogramming. *EMBO J.* **30,** 1778–1789 (2011).
6. Gaspar-Maia, A., Alajem, A., Meshorer, E. & Ramalho-Santos, M. Open chromatin in pluripotency and reprogramming. *Nat. Rev. Mol. Cell. Biol.* **12,** 36–47 (2011).
7. Wang, J. *et al.* A protein interaction network for pluripotency of embryonic stem cells. *Nature* **444,** 364–368 (2006).
8. Yildirim, O. *et al.* Mbd3/NURD complex regulates expression of 5-hydroxymethylcytosine marked genes in embryonic stem cells. *Cell* **147,** 1498–1510 (2011).
9. Sansam, C. G. & Roberts, C. W. Epigenetics and cancer: altered chromatin remodeling via Snf5 loss leads to aberrant cell cycle regulation. *Cell Cycle* **5,** 621–624 (2006).
10. Reisman, D., Glaros, S. & Thompson, E. A. The SWI/SNF complex and cancer. *Oncogene* **28,** 1653–1668 (2009).
11. Kornberg, R. D. Chromatin structure: a repeating unit of histones and DNA. *Science* **184,** 868–871 (1974).
12. Luger, K., Mader, A. W., Richmond, R. K., Sargent, D. F. & Richmond, T. J. Crystal structure of the nucleosome core particle at 2.8A resolution. *Nature* **389,** 251–260 (1997).
13. Li, B., Carey, M. & Workman, J. L. The role of chromatin during transcription. *Cell* **128,** 707–719 (2007).
14. Morse, R. H. Transcription factor access to promoter elements. *J. Cell. Biochem.* **102,** 560–570 (2007).
15. Dennis, J. H. *et al.* Independent and complementary methods for large-scale structural analysis of mammalian chromatin. *Genome Res.* **17,** 928–939 (2007).
16. Li, Z., Schug, J., Tuteja, G., White, P. & Kaestner, K. H. The nucleosome map of the mammalian liver. *Nat. Struct. Mol. Biol.* **18,** 742–746 (2011).
17. Schones, D. E. *et al.* Dynamic regulation of nucleosome positioning in the human genome. *Cell* **132,** 887–898 (2008).
18. Teif, V. B. *et al.* Genome-wide nucleosome positioning during embryonic stem cell development. *Nat. Struct. Mol. Biol.* **19,** 1185–1192 (2012).
19. Valouev, A. *et al.* Determinants of nucleosome organization in primary human cells. *Nature* **474,** 516–520 (2011).
20. Woo, C. J., Kharchenko, P. V., Daheron, L., Park, P. J. & Kingston, R. E. A Region of the Human HOXD Cluster that Confers Polycomb-Group Responsiveness. *Cell* **140,** 99–110 (2010).
21. Li, Z. *et al.* Foxa2 and H2A.Z mediate nucleosome depletion during embryonic stem cell differentiation. *Cell* **151,** 1608–1616 (2012).
22. Tolstorukov, M. Y. *et al.* Swi/Snf chromatin remodeling/tumor suppressor complex establishes nucleosome occupancy at target promoters. *Proc. Natl Acad. Sci. USA* **110,** 10165–10170 (2013).
23. Weiner, A., Hughes, A., Yassour, M., Rando, O. J. & Friedman, N. High-resolution nucleosome mapping reveals transcription-dependent promoter packaging. *Genome Res.* **20,** 90–100 (2010).
24. Henikoff, S., Henikoff, J. G., Sakai, A., Loeb, G. B. & Ahmad, K. Genome-wide profiling of salt fractions maps physical properties of chromatin. *Genome Res.* **19,** 460–469 (2009).
25. Stadtfeld, M., Maherali, N., Borkent, M. & Hochedlinger, K. A reprogrammable mouse strain from gene-targeted embryonic stem cells. *Nat. Methods* **7,** 53–55 (2010).
26. Yuan, G. *et al.* Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* **309,** 626–630 (2005).
27. Segal, E. *et al.* A genomic code for nucleosome positioning. *Nature* **442,** 772–778 (2006).
28. Kharchenko, P. V., Woo, C. J., Tolstorukov, M. Y., Kingston, R. E. & Park, P. J. Nucleosome positioning in human HOX gene clusters. *Genome Res.* **18,** 1554–1561 (2008).
29. Tolstorukov, M. Y. *et al.* Histone variant H2A.Bbd is associated with active transcription and mRNA processing in human cells. *Mol. Cell.* **47,** 596–607 (2012).
30. Henikoff, J. G., Belsky, J. A., Krassovsky, K., MacAlpine, D. M. & Henikoff, S. Epigenome characterization at single base-pair resolution. *Proc. Natl Acad. Sci. USA* **108,** 18318–18323 (2011).
31. Koerber, R. T., Rhee, H. S., Jiang, C. & Pugh, B. F. Interaction of transcriptional regulators with specific nucleosomes across the Saccharomyces genome. *Mol. Cell.* **35,** 889–902 (2009).
32. Gaffney, D. J. *et al.* Controls of nucleosome positioning in the human genome. *PLoS Genet.* **8,** e1003036 (2012).
33. Chung, H. R. *et al.* The effect of micrococcal nuclease digestion on nucleosome positioning data. *PLoS ONE* **5,** e15754 (2010).

34. Johnson, W. E. *et al.* Model-based analysis of tiling-arrays for ChIP-chip. *Proc. Natl Acad. Sci. USA* **103,** 12457–12462 (2006).
35. Cheung, M. S., Down, T. A., Latorre, I. & Ahringer, J. Systematic bias in high-throughput sequencing data and its correction by BEADS. *Nucleic Acids Res.* **39,** e103 (2011).
36. Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res.* **40,** e72 (2012).
37. Kornberg, R. D. & Lorch, Y. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* **98,** 285–294 (1999).
38. Bernstein, B. E. *et al.* A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell* **125,** 315–326 (2006).
39. Deaton, A. M. & Bird, A. CpG islands and the regulation of transcription. *Genes Dev.* **25,** 1010–1022 (2011).
40. Whyte, W. A. *et al.* Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell* **153,** 307–319 (2013).
41. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155,** 934–947 (2013).
42. Chen, K. *et al.* DANPOS: dynamic analysis of nucleosome position and occupancy by sequencing. *Genome Res.* **23,** 341–351 (2013).
43. Bilic, J. & Izpisua Belmonte, J. C. Concise review: induced pluripotent stem cells versus embryonic stem cells: close enough or yet too far apart? *Stem Cells* **30,** 33–41 (2012).
44. Papp, B. & Plath, K. Epigenetics of reprogramming to induced pluripotency. *Cell* **152,** 1324–1343 (2013).
45. Loh, Y. H. *et al.* The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells. *Nat. Genet.* **38,** 431–440 (2006).
46. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133,** 1106–1117 (2008).
47. Chambers, I. & Tomlinson, S. R. The transcriptional foundation of pluripotency. *Development* **136,** 2311–2322 (2009).
48. Matsuda, T. *et al.* STAT3 activation is sufficient to maintain an undifferentiated state of mouse embryonic stem cells. *EMBO J.* **18,** 4261–4269 (1999).
49. Fu, Y., Sinha, M., Peterson, C. L. & Weng, Z. The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. *PLoS Genet.* **4,** e1000138 (2008).
50. Hu, G. *et al.* Regulation of nucleosome landscape and transcription factor targeting at tissue-specific enhancers by BRG1. *Genome Res.* **21,** 1650–1658 (2011).
51. Kornberg, R. D. & Stryer, L. Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res.* **16,** 6677–6690 (1988).
52. Mavrich, T. *et al.* Nucleosome organization in the *Drosophila* genome. *Nature* **453,** 358–362 (2008).
53. Allan, J., Fraser, R. M., Owen-Hughes, T. & Keszenman-Pereyra, D. Micrococcal nuclease does not substantially bias nucleosome mapping. *J. Mol. Biol.* **417,** 152–164 (2012).
54. Soufi, A., Donahue, G. & Zaret, K. S. Facilitators and impediments of the pluripotency reprogramming factors' initial engagement with the genome. *Cell* **151,** 994–1004 (2012).
55. Stadtfeld, M. *et al.* Aberrant silencing of imprinted genes on chromosome 12qF1 in mouse induced pluripotent stem cells. *Nature* **465,** 175–181 (2010).
56. Soufi, A. & Zaret, K. S. Understanding impediments to cellular conversion to pluripotency by assessing the earliest events in ectopic transcription factor binding to the genome. *Cell Cycle* **12,** 1487–1491 (2013).
57. Hawkins, R. D. *et al.* Distinct epigenomic landscapes of pluripotent and lineage-committed human cells. *Cell Stem Cell* **6,** 479–491 (2010).
58. Hu, G. *et al.* H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation. *Cell Stem Cell* **12,** 180–192 (2013).
59. Bryant, G. O. *et al.* Activator control of nucleosome occupancy in activation and repression of transcription. *PLoS Biol.* **6,** 2928–2939 (2008).
60. Zwaka, T. P. & Thomson, J. A. Homologous recombination in human embryonic stem cells. *Nat. Biotechnol.* **21,** 319–321 (2003).
61. Bowman, S. K. *et al.* Multiplexed Illumina sequencing libraries from picogram quantities of DNA. *BMC Genomics* **14,** 466 (2013).
62. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
63. Tolstorukov, M. Y., Kharchenko, P. V., Goldman, J. A., Kingston, R. E. & Park, P. J. Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. *Genome Res.* **19,** 967–977 (2009).
64. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11,** R106 (2010).
65. Machanick, P. & Bailey, T. L. MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* **27,** 1696–1697 (2011).

## Acknowledgements

We thank S. Bowman and M. Simon for optimizing sequencing library preparation, Z. Wang, C. Woo, J. Dennis, and the Kingston and Park labs for helpful discussions, G.Q. Daley for human cell lines, and the MGH Molecular Biology NextGen Sequencing Core. J.A.W., R.E.K., and P.J.P. were supported by the National Institute of General Medical Sciences, NIH (F32GM093491 to J.A.W.; R01GM043901 and R37GM048405 to R.E.K.; R01GM082798 to P.J.P.). K.H. was supported by the grant R01HD058013 from the National Institute of Child Health and Human Development, NIH.

## Author contributions

J.A.W. performed mouse cell line experiments; A.C. performed human cell line experiments and participated in data analysis; B.H.A. helped develop bioinformatic tools; M.S. and K.H. isolated all mouse cell lines and functionally characterized the mouse pluripotent lines; A.D. provided expertise in interpretation of the results, M.Y.T. analyzed the data, J.A.W., A.C., M.Y.T., P.J.P. and R.E.K. designed the study, interpreted the results and wrote the paper. All authors read and contributed to the editing of the manuscript during its preparation.

## Additional information

**Accession codes:** All data sets are available in the NIH GEO database under accession code GSE59064.

**Supplementary Information** accompanies this paper at http://www.nature.com/naturecommunications

**Competing financial interests:** The authors declare no competing financial interests.

**Reprints and permission** information is available online at http://npg.nature.com/reprintsandpermissions/

**How to cite this article:** West, J. A. *et al.* Nucleosomal occupancy changes locally over key regulatory regions during cell differentiation and reprogramming. *Nat. Commun.* 5:4719 doi: 10.1038/ncomms5719 (2014).